Evaluation of Response to Intervention Practices for Elementary School Reading: School and Staff Practices

Supporting Statement for OMB Data Collection Clearance Request

Justification Part B

September 28, 2011
Revised January 19, 2012

Prepared by MDRC for

The Institute of Education Sciences
Contract ED-04-CO-0111/003

PART B: COLLECTION OF INFORMATION EMPLOYING STATISTICAL METHODS

INTRODUCTION

The National Center for Education Evaluation (NCEE) of the Institute of Education Sciences (IES), U. S. Department of Education (ED) is conducting the National Assessment of the Individuals with Disabilities Education Improvement Act of 2004 (IDEA 2004, P.L. 108-446), part of which includes an Evaluation of Response to Intervention (RtI) practices in elementary school reading. Under certain conditions,[1] RtI may qualify as an early intervening service (EIS) that may be supported with IDEA funds to identify and serve students in general education classrooms who may be at risk for academic difficulties and eligible for special education. IES has contracted with MDRC, SRI International, and RG Research Group to conduct the Evaluation of RtI Practices in Reading project. This submission seeks clearance for the data collection instruments and analytical techniques of a study of RtI design, implementation, and impact.

This evaluation is part of the National Assessment of the Individuals with Disabilities Education Improvement Act of 2004 (IDEA 2004, P.L. 108-446) being conducted by IES. Section 664 of IDEA 2004 requires the National Assessment to evaluate "the implementation of programs assisted under this title and the impact of such programs on… improving the academic achievement of children with disabilities to enable the children to reach… challenging State academic content standards based on State academic assessments." MDRC is undertaking the collection of information under contract with IES for this evaluation. This introduction summarizes the study objectives and the three research questions, the specifics of the analytic approach to addressing the three research questions, and data collection plans for the evaluation. This document also provides supporting statements for each of the five points outlined in Part B of the OMB guidelines for the collection of information for the RtI project.

Study Objectives and Research Questions

The RtI approach has the potential to:

1. improve instruction for all struggling students by identifying learning problems early and informing instructional decisions regarding the type, intensity, and duration of interventions to address them;

2. inform the evaluation of students for specific learning disabilities by assessing their responses to research-based interventions; and

[1] Knudsen, 2008.

3. affect the representation of students from culturally and linguistically diverse backgrounds in some disability categories by identifying and intervening early with students who have achievement deficits.

As the study has progressed, it has become clear that there is intense interest in RtI for elementary school reading. As of 2010, 43 states have indicated that they have a state RtI framework in place (retrieved August 22, 2010, from http://state.rti4success.org/). Many districts and schools are working to put in place strong RtI models, and investigation of various types of RtI practices along with quasi-experimental analyses of their impacts can help school district, and state administrators design and implement these programs and inform Federal efforts to support RtI and related early intervening services.

Thus, this evaluation will address the following questions:

1. **What is the average impact on academic achievement of providing intensive secondary reading interventions to elementary school children who have been identified as at risk for reading difficulties compared with children just above the cut point for providing intervention?**

2. **How do academic outcomes, including reading achievement and special education identification, vary with elementary schools' adoption of Response to Intervention practices for early grade reading?**

3. **How do Response to Intervention practices for early grade reading vary across schools and how are they related to academic outcomes?**

The study team will use a regression discontinuity design (RDD) to answer the first question. The RDD analysis will examine the impacts of providing more intensive reading support to children on the margin of needing such assistance. In sites where decisions about providing assistance are made based on a ranking of students' need for assistance and a consistently applied cutoff for assistance, RDD impact estimates would be calculated by comparing student academic outcomes for children immediately above and below the cutoff point. This analysis would provide evidence on the effectiveness of providing coordinated early intervention services (CEIS) funded under IDEA to students who are at the time not identified as needing special education services but are struggling to learn how to read and are receiving more intensive instructional supports for reading in the regular education classroom (Tier 1 in RtI terminology) or in separate tiers with small student-to-teacher ratios.[2]

[2] The Office of Special Education Programs recently issued guidance to provide States with information regarding the use of funds provided under Part B of the Individuals with Disabilities Education Act by local educational agencies (LEAs) to develop and implement coordinated early intervening services (CEIS) for students who are currently not identified as needing special education.

A comparative interrupted time series (CITS) design will be used to answer the second question. The CITS analysis will examine whether implementation of RtI practices is associated with greater improvements over prior academic trends in reading achievement and special education identification in schools experienced with RtI as compared to similar schools not implementing the key elements of reading RtI during the period of the analysis. This design will also examine how special education referral and placement change as RtI is implemented.

For the third question, study team will document the design and implementation of RtI in the full sample of schools (RDD and CITS) through correlational analysis of surveys of school-level staff, teachers, and reading specialists (sometimes known as interventionists). These surveys will also inform the RDD and CITS analyses by allowing us to characterize the contrast in instruction provided students identified as needing additional, intensive reading instruction and those not identified for such services. For the CITS analysis, it will also provide information on the service contrast between the RtI treatment schools and comparison schools.

Research Question #1 Addressed by a Regression Discontinuity Analysis

This approach will compare (1) reading achievement outcomes for students who, based on their benchmark reading test scores, qualified to receive additional reading support with (2) achievement outcomes for students in the same school who meet reading benchmarks initially and were not identified for extra help in reading. Experienced RtI schools typically use a benchmark test at the beginning of the fall semester to identify students for additional reading support. Students whose benchmark test scores fall below a pre-determined cutoff point are deemed at-risk and are referred to additional instructional support (treatment group), and those whose benchmark test scores are above the cutoff stay in the general education class (comparison group). The so-called "sharp" RDD assumes that the decision on receiving the added support is entirely determined by the benchmark test score. The so-called "fuzzy" RDD can accommodate a situation where other factors also influence the decision about receipt of extra support leading to a situation where some students identified for the treatment group based on the benchmark test score do not actually get the extra support and some students identified to receive regular services do get extra support. [3] Therefore, by statistically controlling for the value of the benchmark test score in a regression model, one can (under appropriate conditions) account for any unobserved differences between the treatment and comparison group and thereby obtain internally valid impact estimates for receiving more intensive, additional reading support.

The sample of schools for the RDD analysis will include schools that:

- maintain benchmark test data for each student.

[3] See Van Der Klaauw (2008) and Shadish et al. (2002).

- can provide information about the process of identifying students for additional instructional support, including whether identification involved a decision process based on a single benchmark score, or whether multiple benchmark test scores (and/or other factors) were used to identify students for support.

- assign students to treatment or non-treatment status (i.e., receipt or non-receipt of more intensive reading instruction under a Tier 2 intervention or other means) based on whether their value for a numeric rating (benchmark test score) is above or below a cutoff point;[4]

- maintain a record of the cutoff point(s) used to assign students to receive additional instructional support.

- maintain records tracking students' receipt of extra reading support status throughout the year.

- are willing to allow study-administered year-end reading testing in first and perhaps second grade and can provide spring reading test scores for third graders.

If the above conditions are present and if we can correctly account for the relationship between the benchmark test score and the outcome measure in a statistical model, then this approach can provide an internally valid estimate for the impact on at-risk students' reading achievement of being identified to receive additional instructional support within an RtI system.

Statistical Model. Regression discontinuity analysis was introduced by Thistlethwaite and Campbell (1960) and has more recently experienced a resurgence of interest (e.g., van der Klaauw, 1997, 2002; Angrist and Lavy, 1999; Hahn, Todd, and van der Klaauw, 2001; for a recent review of the literature, see Lee and Lemieux, 2009). This design is the strongest quasi-experimental method that exists for estimating program impacts in the sense that, under certain conditions, this method can approach the rigor of a randomized experiment. In what follows, we describe key elements of the design and analytic methods associated with it.

The regression discontinuity design capitalizes on the systematic process used by experienced RtI schools to identify at-risk students to receive additional reading support within the RtI system. Often, experienced RtI schools use a benchmark test at the beginning of the fall semester to identify at-risk students for additional support. This approach will compare reading achievement outcomes for at-risk students who, based on their benchmark test scores, *just* qualified to receive additional reading support with achievement outcomes for students in the same

[4] A fundamental RDD assumption is that students' ratings and the cut-off point are determined independently of each other – such that assessments of individual students' reading abilities are not influenced by considerations about whether to provide additional support to such students. The study team will verify this assumption's validity during follow-up conversations with experienced RtI schools.

school who *just* meet reading benchmarks initially (likely focusing on students near the cut off for Tier 2 intervention).

In order to attribute the entire shift in outcome at the cut-off point to the treatment one must be able to assume that the following three conditions are met:

- The outcome/benchmark score regression has the correct functional form and is a continuous function throughout the analysis interval absent of the treatment,
- The cut-off point is determined independently of the benchmark score, and
- Nothing other than treatment status is discontinuous in the analysis interval (i.e. there are no abrupt changes in the characteristics of the students included in the interval)

Under these conditions, it is valid to infer that the discontinuity in the outcome-by-benchmark score relationship was caused by the shift in treatment status at the cut-off point. Thus, by statistically controlling for the value of the benchmark test score in a regression model, one can account for any unobserved differences between the treatment and comparison group and thereby obtain internally valid impact estimates for receiving more intensive, additional reading support.[5]

If the decision on receiving the added support is entirely determined by the benchmark test score, in other words, all students whose scores are below the cut-off point would be selected for extra help while all students above the cut-off points would not, then we would have what is referred to as a ***sharp*** regression discontinuity design. In the context of this sharp regression discontinuity, the impact of being identified for extra help can be estimated by the shift in the outcome-by-benchmark score regression at the cut-off point.

Equation 1 provides a simple way to make the regression discontinuity estimation procedure operational for *a single school* in the current study.

$$Y_i = \alpha + \beta Treat_i + \gamma f(R_i) + \sum_{m=1}^{M} \delta_m X_{mi} + \varepsilon_i$$

$$Y_i = \alpha + \beta Treat_i + \gamma f(R_i) + \sum_{m=1}^{M} \delta_m X_{mi} + \varepsilon_i \qquad (1)$$

where

$Y_i Y_i$ = the outcome measure for student i

$Treat_i Treat_i$ = 1 if student i is identified to receive additional instruction and 0 otherwise

[5] For more detailed discussion of the conditions that ensures the design's internal validity, see Lee (2008) and Lemieux (2009).

$f(R_i)f(R_i)$ = a function of the benchmark test score for student i[6]

$X_{mi}X_{mi}$ = the mth background characteristic (for example, race/ethnicity, free and reduced-price lunch status, etc) for student i (m = 1, 2, ..., M)[7]

$\varepsilon_i\varepsilon_i$ = a student level random error term, assumed to be independently and identically distributed

The coefficient for treatment assignment, $\beta\beta$, represents the marginal impact of being identified for extra help at the cut-off point.

Note that this study is composed of separate regression discontinuity designs for each of the RtI schools in the sample. Each school in the RDD sample is considered a "mini" RDD of its own. To estimate the average impact of being identified to receive additional help, the impact estimates for the RD analysis would be pooled across all schools in the RD sample.[8] Specifically, Equation 2 would be used to estimate the *average impact* of being identified to receive additional help for an average student in the RtI schools in the sample.

$$Y_{ij} = \sum_{j=1}^{J} \alpha_j S_j + \beta Treat_{ij} + \sum_{j=1}^{J} \gamma_j f(R_{ij}) * S_j + \sum_{m=1}^{M} \delta_m X_{mij} + \varepsilon_{ij}$$

$$Y_{ij} = \sum_{j=1}^{J} \alpha_j S_j + \beta Treat_{ij} + \sum_{j=1}^{J} \gamma_j f(R_{ij}) * S_j + \sum_{m=1}^{M} \delta_m X_{mij} + \varepsilon_{ij} \qquad (2)$$

where

S_jS_j = 1 if school j and 0 otherwise. This is a dichotomous indicator for school j (j = 1, 2,..., J)
and all other variables are defined as above.

In this model, the impact estimate, $\beta\beta$, is the fixed-effects, average impact across all schools in the RD sample, weighted by the number of students in each school. If the aforementioned conditions are met, and if we can correctly account for the relationship between the benchmark test score and the outcome measure in the model, then the estimated $\beta\beta$ can provide internally valid estimate for the impact

[6] The benchmark test score is usually called the "rating" variable or the "running" variable. The functional form of the relationship between this rating variable and outcome can be estimated by various parametric and/or nonparametric methods have been proposed to estimate the functional form of the relationship between outcome and the rating variable. The difficulty of this task has been cited in the literature as perhaps the most serious limitation of the regression discontinuity approach. For a detailed discussion of these methods and their limitations, see Lee and Lemieux (2009).

[7] These covariates are added to improve the precision of the impact estimates.

[8] Because different schools have different cut-off standard and might use different benchmark test, to be able to pool the analysis across schools, the benchmark test scores need to be standardized. This can be done by centering each student's test score at the cut-off point of his/her school and dividing the centered score by the standard deviation of the score in that school.

on at-risk students' reading achievement of being identified to receive additional instructional support within an RtI system.

This approach also has the following features:

- School fixed-effects are included to account for mean differences in outcome across schools.
- The relationship between outcome and benchmark test score, $f(R_{ij})f(R_{ij})$, is allowed to vary by school to more accurately capture the functional form of the outcome-by-rating regression.
- Additional student characteristics are included in the model to improve precision. Their relationships with the outcome are assumed to be constant across schools.

A separate model would be estimated for each combination of grade levels and each year included in the study sample.

Discussions above assume that a student's treatment status is solely determined by his or her benchmark test score. When that is not true, i.e., when another factor also influences the decision about receipt of extra support, leading to a situation where some students identified for the treatment group based on the benchmark test score do not actually get the extra support (analogous to "no-shows") and some students identified to receive regular services do get extra support (analogous to "cross-overs"),[9] a ***fuzzy*** regression discontinuity design results. In this situation, the *instrumental variable* method can be used to extract the average effect of treatment on students at the cut-off point who receive treatment because they are assigned to it.[10,11]

The analysis based on the RD design helps to answer questions that are directly relevant to an important issue in educational practice. The Office of Special Education Programs of the Department of Education has recently issued a guideline explaining to states how they can use IDEA funds to provide coordinated early intervening services to students not currently identified as needing special education services. Information about the impact of providing more intensive reading support under an RtI framework will be useful in considering the effectiveness of services to students on the margin of needing extra assistance.

Minimum Detectable Effect Size. The statistical precision of an impact estimator reflects its ability to detect true intervention effects when they exist. A common way to represent precision is a minimum detectable effect size (MDES), which is the smallest true effect size that an estimator has a "good chance" of detecting (Bloom, 1995). Discussion in this section presents, under various assumptions, the total

[9] See van der Klaauw (2008) and Shadish et al. (2002).

[10] This subpopulation is often referred to as "compliers" (Angrist, Imbens, and Rubin, 1996).

[11] Hahn, Todd, and van der Klaauw (2001) first formalized this approach. For a more recent discussion on this approach, see Bloom (2010).

number of schools that are required to achieve the precision target of 0.15 units of standard deviation.[12]

Note that the statistical precision analysis reported here uses the standard convention of defining a minimum detectable effect size as the smallest true impact that has an 80 percent chance of being found to be statistically significant (it has 80 percent statistical power) at the 0.05 level of statistical significance for a two-tailed test of the null hypothesis of no effect. For ease of computation, we also assume the multiplier to be 2.8 across all calculations.[13] In addition, we assume that there are 80 students per school per cohort in a given grade, and the student-level response rate is 85 percent. No multiple hypothesis test adjustment is made in current calculations.

The following equation is used to calculate the require sample size for a regression discontinuity design to achieve a target MDES of 0.15:

$$MDES_{RDD} \approx 2.8 * \sqrt{\frac{\left(1-R_{st}^2\right)}{N*RR*J*P\left(1-P\right)*\left(1-R_T^2\right)}}$$

$$MDES_{RDD} \approx 2.8 * \sqrt{\frac{\left(1-R_{st}^2\right)}{N*RR*S*J*P\left(1-P\right)*\left(1-R_T^2\right)}} \qquad (3)$$

where

J = total number of schools,

N = average number of students per grade per school; assumed to be 80, [14]

RR = student level response rate, assumed to be 85%,

S = proportion of subsample of students that are close to the cutoff to be included in the estimation, assumed to be 30% with 15% above the cutoff and 15% below it.

P = the proportion of students in the treatment group in the subsample of students included in the estimation; assumed to be 0.5,[15]

R_{st}^2 R_{st}^2 = student-level explanatory power of covariates; assumed to be 0.4 (Bloom, et al, 2005).

[12] An MDES of 0.15 reflects the current prevailing standard of precision for evaluations funded by the U.S. Department of Education (ED).

[13] This multiplier depends on the number of degrees of freedom available (Bloom, 1995), but for more than about 20 degrees of freedom its value is roughly 2.8. Most of the situations considered in the discussion have more than 20 degrees of freedom.

[14] The average third grade enrollment for elementary schools (defined as schools with no grade higher than the 6th grade) across the U.S. for 2008-2009 school year was 77.9, based on data reported by the Common Core of Data (CCD),

[15] The T/C ratio is assumed to be 1:3 for the full sample because a smaller proportion of students are expected to be selected/identified for additional instructions. However, the same "bandwidth" is often used to select a subsample of treatment and comparison students who are close to the cutoff point to be included in the estimation. Therefore, for this subsample, the T:C ratio is assumed to be 0.5.

$R_T^2 R_T^2$ = proportion of variation in treatment status (T) predicted by the benchmark test score.

A key parameter in this calculation is the proportion of variation in the treatment status predicted by the rating variable (i.e., the benchmark test score), $R_T^2 R_T^2$. This proportion, in turn, depends on how ratings are distributed around the cut-off point (Goldberger, 1972; Bloom et al., 2005; and Schochet, 2008). We assume a normal distribution in our calculation because ratings in this study are likely to be scores on a test and test scores often follow a normal distribution. To compute $R_T^2$for a given distribution of ratings one can generate ratings (r) from a distribution of interest, attach the appropriate value of the treatment indicator (T) to each rating and regress T on r.[16] Doing so yields an R_T^2 of 0.6367 for a balanced normal distribution.[17]

Substituting these values into Equation 3 and one can solve for the total number of schools needed to achieve a minimum detectable effect size of 0.15 or 0.10 for this design. Assuming using a subsample of 30% of the full sample students (15% above the cutoff and 15% below it) that are close to the cutoff, the required number of school is 113 for a targeted MDES of 0.15 and 254 for a targeted MDES of 0.10.

Research Question #2 Addressed by a Comparative Interrupted Times Series Analysis

Under a CITS design, trends in student outcomes such as reading achievement, grade promotion, and special education identification prior to the implementation of RtI practices are compared with post-implementation trends in these schools to estimate a deviation from prior trends occurring with the start of RtI (the "interruption"). This deviation in RtI treatment schools is then compared with an estimated deviation in outcomes that occurred in similar schools not implementing RtI across the same period. The estimated difference in these two deviations is the estimate of the "impact" or more properly the "association" between the adoption of RtI practices and student outcomes. The causal evidence emerging from this methodology is weaker than for either regression discontinuity or random assignment studies.

The sample of schools for the CITS analysis will have the following characteristics:

- Sufficient numbers of "treatment" schools that have experience with RtI practices to have the needed statistical power to detect relationships (as discussed elsewhere in this submission);
- Experienced RtI schools that have good historical information about the timing of RtI implementation;

[16] Note that here we are also assuming the functional form to be linear, as test scores tend to be.

[17] This approach is also used in Bloom (2010) to estimate sample size multiples for RD design.

- Experienced RtI schools that are implementing RtI practices with a clearly identifiable starting point;

- Appropriate, statistically-equivalent comparison schools that can be systematically identified;

- Treatment and comparison schools that have historical data on student outcomes measured using consistent metrics over three or more years[18] prior to the first year of RtI implementation in the experienced RtI schools; and

- Treatment and comparison schools that have one or more years of follow-up data, measured using the same metrics as those used for the historical data, in the period following RtI implementation in the experienced RtI schools.

In this analysis, we will collect existing student records for special education referral and identification and disability category and – as available - reading achievement during the baseline period prior to RtI implementation (the interruption) and in a post-interruption follow up period. Similar data will be collected in RtI treatment schools and matched comparison schools ideally in the same districts as treatment schools. The details of the CITS approach are described below, but we recognize it has less methodological strength in identifying causal relationships. Specifically, it does not provide causal estimates of the impact of implementing RtI practices on the student outcomes examined. Hence, we at times in this submission use the phrase "association between RtI implementation and changes in student outcomes."

In the sections below outlining the statistics of the method, we follow the usual convention in the CITS literature of using the terminology of impacts. The use of the association terminology would complicate the explanation of the approach. In presenting findings from the analysis in the project report, we will take care to signal the weaker causal inferences that must be drawn as compared to the RDD analysis.

Statistical Model. In principle, the impact of RtI on a student outcome equals the *difference* between what the outcome was after RtI was under way and what it would have been without RtI. In the CITS design, an estimate without strong causal inference is calculated by comparing the change over time in a student outcome for schools that adopted RtI (program schools) with the corresponding change for similar comparison schools that did not adopt it (the "counterfactual").[19] Thus, the

[18] The literature does not provide much guidance on the minimum number of baseline years needed. MDRC tends to use three years as a minimum requirement for CITS. In general, longer baseline periods yield better estimates of trends, and, accordingly, yield better estimates of impacts based on deviations from trends.

[19] For a description of this approach—which is referred to as "short interrupted time-series analysis"—see Bloom (2003).

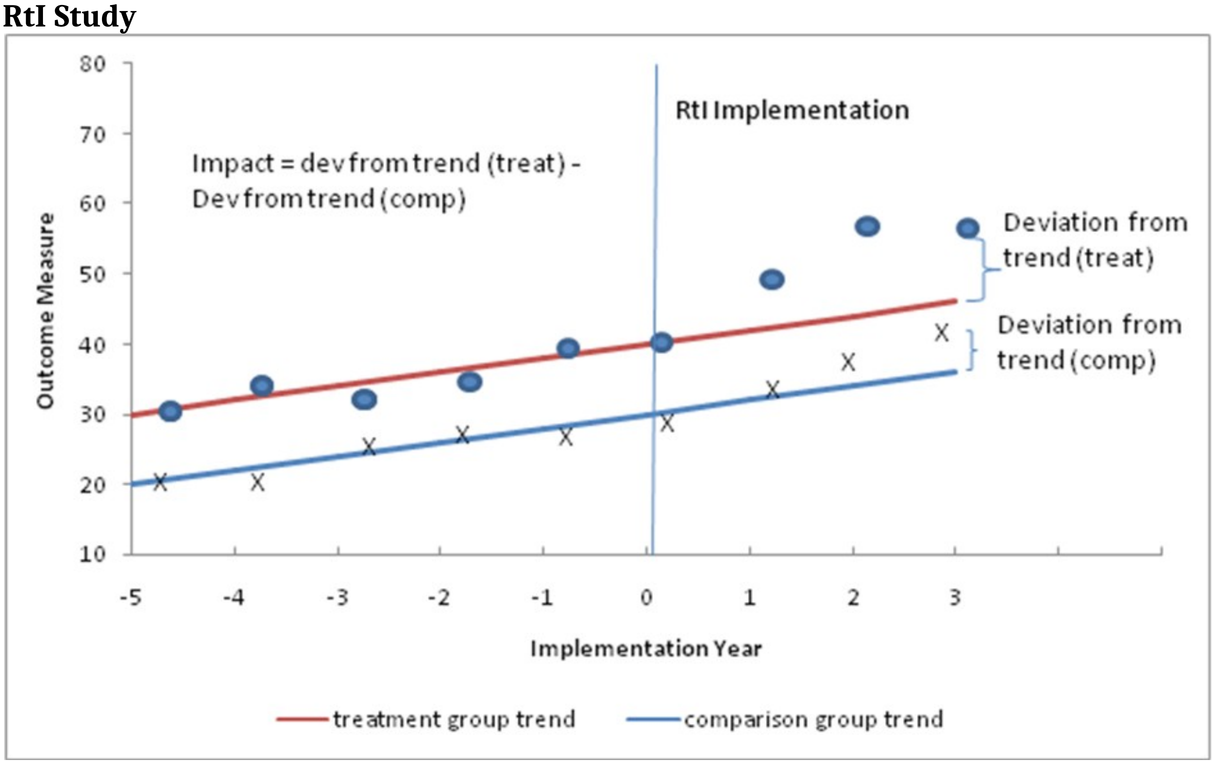
estimate represents the observed improvement of the RtI program schools relative to the observed improvement of their comparison schools. In what follows, we further describe the basic concept of this approach, present the statistical model that will be used for the impact estimation, and discuss the types of data needed for this analysis. Again, we recognize that this method does not produce causal estimates of the impact of RtI pratices, instead providing evidence of the association of implementation of RtI practices with changes in student outcomes.

Ideally the time-series design used to produce estimates should have data on consistently measured student outcomes for multiple pre-intervention baseline years, multiple post-intervention follow-up years, multiple program schools, and multiple comparison schools.[20]

The application of a CITS design to the RtI study is illustrated in the Figure B-1, below. In the first instance, an outcome measure in the program schools (schools that started implementing RtI strategies at the beginning of the follow-up period) at follow-up is compared with the outcome that one would have expected at these schools given their historical patterns/trends in the given outcome (we call this "deviations from trend"). In the second step, deviations from trend in the treatment schools are compared to deviations from trend in a group of comparison schools (schools that do not implement RtI) in order to "subtract out" the effect of common policies that may have been implemented concurrent with the implementation of RtI. The key measure of how student outcome vary with the implementation of RtI is the *difference* between the average deviation from trend in the treatment schools and the average deviation from trend in the comparison schools.

[20] Multiple baseline years help to provide a reliable benchmark and trend of pre-intervention outcome. Multiple follow-up years help to provide the elapsed time needed for a reform to be implemented and thus to begin to take effect. Multiple program schools help to provide a reliable measure of change over time in the presence of the program. This reliability stems from (1) the ability of multi-school averages to reduce random year-to-year fluctuations in student outcomes and (2) their ability to "dampen the shocks" that can occur at a single school due to idiosyncratic local events, such as a change in principal. For the same reasons, multiple comparison schools can help to provide a reliable basis for estimating the change over time in student outcomes that would have occurred without the program.

Figure B-1: Comparative Interrupted Time Series (CITS) Design Applied to the RtI Study



The model used to estimate this "difference in deviation from trend" is presented below. The following discussion focuses on the statistical model used for students' reading achievement test scores, but similar models can be used for other outcomes such as special education identification.

Equations 4 and 5 represent the two-level hierarchical model that will be used in the estimate for student reading test scores for 3 follow-up years.[21] These estimates are based on student outcome data for a given test and for a given grade (for example, the third grade state reading test score or the second grade ORF score) during the baseline and follow-up years for both the program schools and their comparison schools, plus data on student background characteristics.

<u>Level 1: Students within schools</u>

$$Y_{ijt} = \alpha_{0jt} + \sum_{m=1}^{M} \alpha_m X_{mijt} + \varepsilon_{ijt} Y_{ijt} = \alpha_{0jt} + \sum_{m=1}^{M} \alpha_m X_{mijt} + \varepsilon_{ijt} \qquad (4)$$

<u>Level 2: Schools</u>

$$\alpha_{0jt} = \sum_{j=1}^{J} \beta_j S_j + \sum_{j=1}^{J} \gamma_j TIME_t * S_j + \theta_1 FY1_t + \theta_2 FY2_t + \theta_3 FY3_t$$

[21] The model can be adjusted to estimate impacts for more or less than 3 follow-up years.

$$+\delta_1 FY1_t * RtI_j + \delta_2 FY1_t * RtI_j + \delta_3 FY1_t * RtI_j + \mu_{jt}$$

$$+\delta_1 FY1_t * RtI_j + \delta_2 FY1_t * RtI_j + \delta_3 FY1_t * RtI_j + \mu_{jt} \qquad (5)$$

where

$Y_{ijt}Y_{ijt}$ = the outcome for student i in school j from school year t
$X_{mijt}X_{mijt}$ = the mth background characteristic (for example, race/ethnicity, free and reduced-price lunch status, etc) for student i in school j from school year t (m = 1, 2, …, M)
S_jS_j = 1 if school j and 0 otherwise. This is a dichotomous indicator for school j (j = 1, 2,…, J)
$TIME_tTIME_t$ = a continuous variable for relative school year for school j and school year t. The first RtI implementation year is coded as 0, and all other years are coded relative to this value
$FY1_tFY1_t$ = 1 if school year t is the first follow-up year for school j
$FY2_tFY2_t$ = 1 if school year t is the second follow-up year for school j
$FY3_tFY3_t$ = 1 if school year t is the third follow-up year for school j
RtI_jRtI_j = 1 if school j is an RtI program school

Level 1 of the model specifies that the outcome, $Y_{ijt}Y_{ijt}$, for a given student from a given school in a given year depends on his or her background characteristics, $X_{mijt}X_{mijt}$(m = 1, 2, …, M), plus a random error, $\varepsilon_{ijt}\varepsilon_{ijt}$, which is independently and identically distributed. For simplicity, the relationship between each background characteristic and the student outcome (measured by the regression coefficient, $\alpha_m\alpha_m$) is assumed to be constant across schools and time.

Level 2 of the model specifies that the "regression-adjusted mean outcome, $\alpha_{0jt}\alpha_{0jt}$, for all students from a particular school in a particular year depends on the school involved, S_jS_j, the year relative to the first RtI implementation year, $TIME_tTIME_t$, whether the school is a program school, RtI_jRtI_j, whether the year is a follow-up year, and a random error, $\mu_{jt}\mu_{jt}$, which is independently and identically distributed.

Therefore, the coefficient $\beta_j\beta_j$in this model represents the regression-adjusted mean outcome for the baseline period for school j; the coefficient $\gamma_j\gamma_j$ represents the slope of the baseline trend for school j; the coefficient $\theta_n\theta_n$ (n = 1, 2, 3) represents deviation from baseline trend for comparison schools for follow-up year n; and the coefficient $\delta_n\delta_n$ (n = 1, 2, 3) represents the deviation from baseline trend for the program schools that is above and beyond that for comparison schools for follow-up year n, i.e., it represents the *difference* between the average deviation from trend in the RtI schools and the average deviation from trend in the comparison schools.

These latter coefficients are the central ones for the CITS analysis. Note that these are fixed-effects estimates for the average student in the average school in the study sample and should not be generalized to some larger population of schools. This is because the schools in this analysis will be selected purposefully and are not a random sample of schools from a larger target population.

This analytic approach has the following additional features:

1) By controlling for individual students' background characteristics, the statistic model controls for possible compositional shifts over time in the student population.
2) The model controls for unobserved school effects by using a fixed-effects approach: the S_jS_j indicators in Level 2 of the model control for unobserved school effects and therefore the clustering of data within schools.
3) Because observations in a given year are more similar to each other than observations in different years, schools are nested within year, and this clustering structure needs to be accounted for in the estimation. This model deals with time effects by using a random-effects approach, which allows the effect of time to vary randomly across schools in the error term.[22]

Minimum Detectable Effect Size. The following equation is used to calculate the minimum detectable effect size given a CITS design.[23] It states that for a CITS design with cohort differences (i.e., the intra-class correlation (ICC) across cohorts of students is not zero), the MDES can be calculated as the following:[24]

$$MDES_{CITS} \approx 2.8 * \sqrt{\frac{1}{P(1-P)J}}\sqrt{\frac{1}{N*RR}+\frac{ICC}{1-ICC}}\sqrt{1+\frac{1}{T}+\frac{(T_f-\bar{t})^2}{\sum_k (t_k-\bar{t})^2}} \tag{6}$$

where
J J = total number of schools,
P = proportion of treatment schools in the sample; assumed to be 0.5 or 0.33,
N N N = average number of students per school; assumed to be 80,
RR = student level response rate; assumed to be 85%,
ICC = intra-class correlation across student cohorts; assumed to be 0.006,[25]
T = number of baseline years, assumed to be 3,

[22] This can be accomplished by specifying time indicator in the RANDOM statement using PROC MIXED in SAS.

[23] This is based on Equation 4 in Bloom (1999).

[24] This calculation assumes equal number of treatment and comparison schools.

T_f =the follow-up year of interest; takes the value of 0, 1, 2 for follow-up year 1, 2, 3,

$\bar{t}$ = the mean baseline year; takes on the value of -2 for 3 baseline years,

$\sum_k (t_k - \bar{t})^2$ $\sum_k (t_k - \bar{t})^2$ = the sum of squared variation of the baseline years around the mean baseline year; takes on the value of 2 for 3 baseline years.

Based on this calculation, it can be demonstrated that

- The statistical power of a CITS design improves greatly as the number of baseline years included in the estimation model increases because more baseline year data can help estimate and predict the time-trend more precisely.
- The magnitude of the intra-class correlation across student cohorts makes a big difference in the statistical power. In general, smaller ICC provides more precision because smaller ICC indicates less random cohort variations and consequently a more stable time trend.
- More schools are required to achieve the same precision target for later follow-up years because baseline year information is used to estimate and predict the trend for follow-up years, and as the time difference between the follow-up year of interest and the baseline year grows, the precision of the predicted time-trend decreases.

Based on these assumptions, it can be calculated that, to achieve the targeted MDES of 0.15 standard deviations in the first follow-up year, a total number of 96 schools are required for a CITS design with equal number of treatment and comparison schools (i.e., 48 treatment schools and 48 comparison schools). This sample size allows us to achieve an MDES of 0.22 standard deviations and 0.28 standard deviations for the second and third follow-up years, respectively.[26]

In reality, sometimes it might be hard to recruit the same number of treatment and comparison schools for a study, and sometimes one might want to build a buffer zone against any unforeseeable future events (such as school closing) that might

[25] Empirical estimations of this parameter, especially for student academic outcomes in a school setting, are rarely reported in the literature. Bloom (2001) reports a 0.002 value for this parameter using three years of standardized reading scores for third-graders from eight elementary schools located in seven states which adopted an early version of the Accelerated School model in the mid-1990s (Bloom, 2001). Using four year third-grade reading data from 25 elementary schools in Rochester, New York, Bloom (1999) reports that the 75th percentile value of the ICC distribution is 0.01. Both of these two values reflect the possible range of this parameter in a setting that would be similar to the current study. The average of these two numbers, 0.006, is used in calculations reported here.

[26] Note that later follow-up years provide a better chance to see bigger effects since the maturation of the RtI models implemented in the treatment schools.

cause us to drop comparison schools from the sample..[27] Therefore similar calculations were done for an unbalanced CITS design where the T/C ratio is 1:2. In general, for a fixed sample size, any deviation from a balanced design will reduce the statistical power of a design. Therefore it will require a bigger sample to achieve the same level of precision. To achieve a MDES of 0.15 in this design, a total of about 108 schools (i.e., 36 treatment schools and 72 comparison schools) are required.
Barring any attrition of schools later on, a sample of this size can achieve a MDES of 0.21 in the second follow-up year and 0.27 in the third follow-up year.

The study team has planned to conduct student-level subgroup analyses. Based on the sample size calculation results reported above for an MDES of 0.15*, in the first follow-up year*, a student subgroup that consists of 25% of the students in a given grade (e.g., for a certain race/ethnicity group or for students whose prior test score is in the lowest quartile of the score distribution) can achieve a MDES of 0.27 if the T/C ratio is 1:1 and 0.28 if the T/C ratio is 1:2. If the targeted MDES for the full sample analysis is set to be 0.10, a subsample of 25% of the students can achieve a MDES of 0.18 if the T/C ratio is 1:1 and 0.19 if the T/C ratio is 1:2 for the first follow-up year.

Research Question #3 Addressed by a Comparison of Description Statistics

The descriptive analysis of RtI design and implementation will include 3 main elements:

- ***For all study (RtI treatment and comparison) schools:*** A description of the structure of RtI or other programs to assist students in reading, including universal screening or benchmark testing, offerings of reading instruction (whether offered in the general education program (often called Tier 1 in an RtI program) or in more intensive ways (often offered in a second and third tier within an RtI program); progress monitoring of students over time; use of data to make decisions about tier placements and movements; and the extent to which they have a process for determining eligibility for special education services that includes data on student's responsiveness to the interventions.

- ***RDD Treatment Schools:*** Details of the assessment process used for benchmark testing, how these benchmark tests are used in decisions to offer or discontinue more intensive reading support and student receipt of more intensive reading support throughout the school year.

[27] Holding the total number of schools equal, a sample with equal number of treatment and comparison schools has more power; on the other hand, identifying more than one comparison schools for one treatment school provides cushion for potential attrition of schools in the future. The choice between the two to a large degree depends on what's available as the comparison pool in reality.

- ***CITS RtI Treatment and Comparison Schools:*** Details on the timing of the implementation of elements of any RtI program and documentation of the service contrast (presences or absence of RtI practices) over the time of the analysis.

Summary of Site Recruitment

The site recruitment screening materials and process, already approved by OMB under collection 1850-0872, are being used to identify and recruit schools that can be part of these three components of the study. Site recruitment – though not completed – has demonstrated the feasibility of all three components of the study.

The site recruitment process began with us seeking nominations from RtI experts of districts and schools experienced with implementation of RtI. We contacted RtI experts who represent different stakeholders and perspectives in RtI, including researchers[28], practitioners[29], and representatives from organizations supporting RtI activities[30]. Nominators were sent a letter and description of the study that outlined the nomination process and how the information they provided would be used. We have proceeded with recruiting districts and schools that were nominated using the OMB approved screening protocols.

Since the approval of site recruitment materials for the study, the project team has been soliciting nominations and screening potential sites for possible inclusion in the study. We anticipate that this process will continue through the summer of 2011, leading to final study site recommendations to IES in the winter of 2011-2012, to allow for the start of data collection when the OMB package is approved.

While the site recruitment process continues, we have at the time of this submission assembled a pool of potential study sites that demonstrates the feasibility of both quasi-experimental methods for analyzing the impacts of RtI. The study team has focused its recruitment activities in 13 where there are multiple districts nominated to save on project resources for screening and data collection. In these states, 311 schools have been nominated in 162 different districts.

As of December 29, 2011, we have identified 107 candidate schools that satisfy our screening protocols for both the RDD and CITS designs, and an additional 67 schools that satisfy our screening protocol only for the RDD design. We will be completing the screening of some remaining nominated schools, work with IES to identify the schools that are the best match for the requirements of the study, and in the winter of 2011-2012 finalize the selection of sites. We anticipate that some of the

[28] Researchers actively involved in RtI reading were contacted.

[29] Practitioners who have worked extensively with RtI on the ground level – and whose work is recognized by OSEP and IES – were contacted.

[30] Representatives from national organizations working to advance the design and implementation of RtI were contacted.

identified sites will prove not to be a good match with the needs of the study because of the details of how they identify students for intensive reading assistance and, hence, we are seeking a pool of potential sites larger than actually needed.

RESPONSES TO SPECIFIC QUESTIONS

This submission of Part B provides information on: (1) the goals of selection of schools and districts for this study, the methods used to recruit our sample of schools, and the current pool of schools identified as appropriate and interested in participating in the project; (2) our proposed information collection procedures for the analysis of RtI implementation and impacts; (3) methods we have used to maximize response rates in data collection; (4) tests of procedures to be undertaken; and (5) individuals consulted on the statistical aspects of the design.

B1. Respondent Universe and Sampling Methods

The goal of this study is to describe a range of RtI practices to inform practices in the field, to estimate the impact of use of RtI practices to identify and provide students with intensive, secondary reading instruction, and to understand the association between adoption of RtI practices and changes in special education identification rates and student reading achievement. Given these study goals, we have not sought a sample that is statistically representative of all schools but have recruited a sample that includes schools and districts in a diversity of settings and using a variety of RtI practices. Additionally, efforts have been made to recruit larger school districts that contain multiple elementary schools because the clustering of schools in one central location will help to reduce data collection costs and facilitate the analyses by providing a pool of potential comparison schools.

Our evaluation of RtI Practices in Reading has three main components, as discussed in Part A of this submission, linked to our site recruitment process:

- A regression discontinuity design (RDD) analysis, involving 113 treatment schools, 65 used solely for the RDD analysis and 48 treatment schools included in the CITS that are also appropriate for the RDD analysis;
- A comparative interrupted time series (CITS), involving 48 “treatment” schools experienced operating RtI programs and 67 comparison schools in the same or similar districts; and
- A descriptive comparison study including all treatment and comparison schools examining the reading services offered to all students and to those students not meeting local benchmark standards.

In the introduction to this section, we described the characteristics of schools that are appropriate for the two quasi-experimental methods to be used in the analysis. Here we briefly summarize process for site recruitment and our current pool of potential study sites, which illustrates the feasibility of the two quasi-experimental impact designs. The details of these two quasi-experimental designs are described more fully in an earlier later section of this submission.

The site recruitment process began with us seeking nominations from RtI experts of districts and schools experienced with implementation of RtI. We contacted RtI experts who represent different stakeholders and perspectives in RtI, including researchers[31], practitioners[32], and representatives from organizations supporting RtI activities[33]. Nominators were sent a letter and description of the study that outlined the nomination process and how the information they provided would be used.

We have proceeded with recruiting districts and schools that were nominated using the OMB approved Screening protocols. This process has been broken up into five steps to organize our efforts:

> **Step 1**
> *Informational Letter to be Sent to District Directors/Coordinators of RtI or Special Education Services to Identify Experienced Schools* – This letter, sent as an email to nominated districts or districts with one or more nominated schools, introduces the RtI study, explains its purpose, and invites the district to participate in a follow-up phone call to help us learn more about district-wide RtI policies and practices and discuss the prospect of the district's participation in the study.
>
> **Step 2**
> *District Protocol for Obtaining District Information about Experienced RtI Schools and Determining Interest in Participation in the Study* – This is a phone call with someone familiar with RtI at the district level. It defines what is meant by an experienced RtI site, asks the district representative to identify any schools that would fall into the experienced RtI category and any schools that would not (to serve as possible comparison sites), and discusses any district RtI supports available to schools.
>
> **Step 3**
> *School-Level Screening for Identifying Experienced RtI Schools* –This phone call is with the person(s) at the school site who is most familiar with the school's RtI model. It inquires about 1st grade practices related to RtI implementation.

This information is used to help determine whether a site would be a good match for the study as either an experienced RtI treatment school or a comparison site and which design the site would be most appropriate for (RDD, CITS or both). After completing this call, we ask sites if they would be interested in filling out a follow-up form.

Since the approval of site recruitment materials for the study, the project team has

[31] Researchers actively involved in RtI reading were contacted.

[32] Practitioners who have worked extensively with RtI on the ground level – and whose work is recognized by OSEP and IES – were contacted.

[33] Representatives from national organizations working to advance the design and implementation of RtI were contacted.

been soliciting nominations and screening potential sites for possible inclusion in the study. We anticipate that this process will continue through the summer of 2011, leading to final study site recommendations to IES in the winter of 2011-2012, to allow for the start of data collection when the OMB package is approved.

While the site recruitment process continues, we have at the time of this submission assembled a pool of potential study sites that demonstrates the feasibility of both quasi-experimental methods for analyzing the impacts of RtI.

The study team has focused its recruitment activities in 13 where there are multiple districts nominated to save on project resources for screening and data collection. In these states, 311 schools have been nominated in 162 different districts.

There are districts and schools the study team is not pursuing. This is because some districts or schools are not interested in participating in the project (often because they are too busy) and some districts are too new to RtI to be eligible for our study.

Our goal is to identify 113 schools that can serve as RtI treatment schools in the RDD analysis, with at least 48 of these schools also being appropriate as treatment schools for the CITS design. In addition, once we have decided on the final list of CITS treatment schools we will seek comparison schools in the same or similar districts to be included in the study.

As of December 29, 2011, we have identified 107 candidate schools that satisfy our screening protocols for both the RDD and CITS designs, and an additional 67 schools that satisfy our screening protocol only for the RDD design. We will be completing the screening of some remaining nominated schools, work with IES to identify the schools that are the best match for the requirements of the study, and – by the winter of 2011-2012 – finalize the selection of sites. We anticipate that some of the identified sites will prove not to be a good match with the needs of the study because of the details of how they identify students for intensive reading assistance and, hence, we are seeking a pool of potential sites larger than actually needed.

B2. Information Collection Procedures

This submission includes three types of data collection and analysis: (1) estimating the impact of more intensive secondary reading assistance on reading achievement; (2) assessing whether special education identification and – if feasible – reading achievement vary with the adoption of RtI practices for early grade reading; and (3) describing the variation in RtI practices for early grade reading across the study schools. The details of the analytical methods for addressing these topics and the data collected is described in the introduction to this document. Here, we summarize the data to be collected.

Question #1: Impact of intensive, secondary reading instruction on reading achievement addressed using the RDD.

As described in the introduction to this document, this analysis will collect data on fall reading benchmark tests for students in grades 1-3 and decisions based on this regarding the provision of additional, more intensive reading instruction. We will then collect data on reading support services throughout school year 2011-12 and spring 2012 tests scores for reading achievement either from a specially fielded reading test for first and second grade or from existing test data for third grade. This data collection is described in detail elsewhere in this submission. Using these data, we will calculate the impact of intensive, secondary reading instruction on reading achievement using the RDD.

Question #2: Variation in reading achievement and special education identification with adoption of RtI practices, using the CITS Design

In this analysis, we will collect existing student records for special education referral and identification and disability category and – as available - reading achievement during the baseline period prior to RtI implementation (the interruption) and in a post-interruption follow up period. Similar data will be collected in RtI treatment schools and matched comparison schools ideally in the same districts as treatment schools. The details of the CITS approach are described below, but we recognize it has less methodological strength in identifying causal relationships. Specifically, it does not provide causal estimates of the impact of implementing RtI practices on the student outcomes examined.

Question #3: Comparative Description of RtI Design and Implementation

The descriptive analysis of RtI design and implementation will include 3 main elements:

- ***For all study (RtI treatment and comparison) schools:*** A description of the structure of RtI or other programs to assist students in reading, including universal screening or benchmark testing, offerings of reading instruction (whether offered in the general education program (often called Tier 1 in an RtI program) or in more intensive ways (often offered in a second and third tier within an RtI program); progress monitoring of students over time; use of data to make decisions about tier placements and movements; and the extent to which they have a process for determining eligibility for special education services that includes data on student's responsiveness to the interventions.

- ***RDD Treatment Schools:*** Details of the assessment process used for benchmark testing, how these benchmark tests are used in decisions to offer or discontinue more intensive reading support and student receipt of more intensive reading support throughout the school year.

- ***CITS RtI Treatment and Comparison Schools:*** For treatment schools, details on the timing of the implementation of elements of the RtI program. For comparison schools, documentation of the service contrast (presences or absence of RtI practices) over the time of the analysis.

Appendices to this submission include data collection protocols for all three components of the study.

B3. Methods to Maximize Response Rates

The target response rate for information obtained through this collection is 85 percent. The research team will work to establish strong partnerships with the participating districts and schools. These partnerships will rely on effective communication with, and monitoring of, the districts and schools. Constant communication will allow potential concerns to be addressed by all parties and allow the study team to monitor attrition of districts or schools from the project and lessen the chance that selected school and district staff are not participating in the field research activities. If the study team determines that a school or district has opted out of the study, the team will work with the district or school to determine the source of this decision and see if study participation can be achieved.

Teachers who complete study data collection activities (teacher survey or interventionist survey) will receive a gift certificate at a local book or school supply store of $25 for each instrument. Teachers who complete the description of student's reading instruction and intervention will receive a $10 gift certificate for each student per wave of data collection.

Justification for Respondent Incentives. The teacher and interventionist surveys that will be used to collect data from teachers in the RtI treatment and comparison schools have some unique qualities that make administration difficult and lead us to request compensation to assure the needed high response rates. Aspects of the survey effort that may make it more difficult to obtain high completion rates are:

- The surveys for teachers and interventionists are asking about complicated material and hence are fairly lengthy and will require the careful attention of respondents. The surveys are also the only feasible source of information on reading instruction and intervention provided to students in the study schools so a high response rate is important.

- In most cases, the teachers and interventionists will be filling out these surveys on their own time, rather than during the school day when they are paid but have teaching responsibilities. If we were to ask the teachers to complete this survey during the school day, the school would need to provide substitute teachers to cover their classes and would be more costly for the study.

- The instructional logs for our sample of children in each treatment school require the regular grade level teacher to collect information from other teachers if the child is receiving reading interventions beyond that provided by the grade level teacher in the core reading instruction. Therefore, this constitutes additional effort on the part of the grade level teacher to coordinate with additional staff and extra effort on the part of these staff to provide the information. We anticipate that about half of the sampled children will be receiving this additional reading instruction.

- Unlike many IES projects, the schools and staff in the study are not receiving any concrete benefits from participating in the project, such as training or materials. In this study, we have identified schools already implementing RtI practices and have identified other schools that can serve as comparison schools because they are not operating RtI. We want to collect data from both the treatment and comparison schools. Since participation in the project does not already bring benefits to the school, it is especially important to compensate respondents for the effort involved in participating in the data collection activities to obtain the necessary response rates.

- We are not asking for compensation for the principals or RtI coordinator, as unlike the teachers, they have time during their school day to complete the school survey.

These difficulties interact to make these surveys of the school’s teaching staff more difficult to conduct than many surveys in IES projects.

Thus, we are requesting clearance to use respondent payments for those who complete the teacher and interventionist surveys to obtain completion rates that will yield credible results, to avoid the bias that could result from selective non-response, and to reduce item non-response. We believe that the previous experiences and studies of the issue of the effects of compensation on response rates make a strong case for the use of compensation for completing this study’s data collection instruments.

Amount of Proposed Incentives. To be effective, the amount of the incentives must fit the burden of the survey. We have based the amount to be paid to respondents on prior research and on the time burden and estimated hourly compensation of staff. We propose a $25 compensation for the teacher survey (approximately 45 minutes to complete) and the interventionist survey (approximately 30 minutes to complete). Using the incentive chart below, each qualifies as a high burden activity and could receive a payment of up to $30.

Response incentives[33] For surveys, low burden = 10 minute survey of basic background, classroom or school characteristics; medium burden = 20 minute survey of classroom or parental practice or school environment; high burden = 30 minute survey of detailed information on instructional practice, school-level interventions, or parent/student histories and experiences. For teacher assessments, low burden = classroom observations, medium burden = 30 minute survey of teacher knowledge and skills, high burden = formal assessment of teacher knowledge and skills with normed test. For student assessments, low burden = individually scheduled assessment in school; medium burden = student must travel to test administration site; high burden = student and parent must travel to test administration site.

	Low Burden	Medium Burden	High Burden
Teacher or principal survey	$10	$20	$30
Teacher assessment	25	50	100
Teacher rating of students	3 per student	5 per student	10 per student
Parent or student survey/interview	15	25	50
Student assessment	50	75	100

We are also requesting a payment for teachers and any interventionists completing the instructional logs for our sample of students. To meet the RDD requirements and criteria, we propose to draw a sample of eight students in grades 1, 2 and 3 distributed around the cutpoint on the fall benchmark for identifying the students requiring intensive reading interventions. We anticipate that these eight students will be taught by an average of 2.5-3.0 teachers in a school for a total of approximately eight teachers in each school. For these eight students, grade level teachers and intervention providers will report more detailed data on the reading instruction and intervention the sampled students receive during school year 2011-12. This log will be completed for five consecutive days of instruction, up to three times a year, depending upon the final date of OMB approval and the school schedule.. Our burden estimate is that each teacher involved in this data collection will spend 30 minutes per wave providing information on her own reading instruction and any involved interventionists will spend a similar amount of time per wave. According to the IES burden chart, this would support payment of $30 per wave as compensation. Our estimate is that each teacher would be providing information on an average of approximately 2.5 to 3 students and we are requesting compensation of $10 per student per wave, consistent with the guidelines above.

Research Support for Incentive Payments. The best statement of research assessing the use of incentives is the Symposium on Providing Incentives to Survey Respondents convened in October 1992 by the Council of Professional Associations on Federal Statistics (COPAFS) for OMB and a follow up seminar with multiple papers organized by CPAFS in 2008. In 1992, COPAFS asked Richard Kulka of NORC to write a review of the literature in light of what was learned at the symposium. Kulka concluded, "the greatest potential effectiveness of monetary incentives appears to be in surveys that place unusual demands upon the respondent, require continued cooperation over an extended period of time, or when the positive forces on respondents to cooperate are fairly low." Kulka also wrote, "there is evidence that increasing the size of a monetary incentive will result in increases in survey response and/or response quality, although there is also consistent evidence that this benefit may rather quickly reach 'diminishing returns', whereby large incentives no longer result in appreciable increases in survey response" (Kulka, 1992). In more recent work, Kulka has continued to find incentives useful in increasing response rates and response quality and explored in more detail the best ways to structure incentives and the situations in which they are appropriate and useful.

Earlier studies have shown that when used appropriately, incentives are a cost-effective means of significantly increasing response rates (e.g., Dillman, 1978; James and Bolstein, 1990). As Groves, Cialdini, and Couper (1992) note, people feel obligated to reward positive behavior (such as being provided with an incentive) with positive behavior in return—in the current context, such positive return behavior would be defined as a completed survey. Surveys that use incentives can actually be less expensive than those that do not. Respondent incentives can substantially increase cooperation rates and may make the survey less expensive if they result in less need for callbacks or lower missing-data rates.

We believe that the studies summarized here, and the study team's previous experiences with fielding surveys and other kinds of assessments, make a strong case for the use of respondent payments for completing the teacher surveys in this study

B4. Tests of Procedures to be Undertaken

The study team has developed a data collection plan that is designed to produce high quality data for the study.

Analysis of Variation in RtI Practices through Survey Data Collection

For the implementation research and case studies, the study protocols include items that have been used in prior studies of reading intervention programs or that were piloted by the study team in the spring of 2010 in five schools (involving under 10 respondents). This led to substantial revisions to the instruments, including streamlining of many instruments. School and district staff were very cooperative during these pilot visits and we believe we can sustain this level of cooperation in the full scale data collection.

Measures of Student Reading Achievement for the RDD

In this and the following section, we describe the planned reading achievement outcome measures to be used in this study. We first address the measures for the regression discontinuity design (RDD) and then discuss the reading measures used for the comparative interrupted time series (CITS) design. For each of the designs, we identify a set of criteria for selecting outcome measures and provide a rationale for why these criteria are important. Since two of the criteria are identical for both designs, we begin with these.

General Criteria for Selection of Reading Achievement Measures

1. ***The measure needs to address reading proficiency in a broad, reasonably comprehensive fashion.***

Rationale: The goal of RtI is to improve reading proficiency for students at risk for failure. RtI can focus on any one or more aspects of reading including oral reading, silent reading, phonics, fluency, comprehension, vocabulary– depending on student need and district philosophy. Therefore, an evaluation should encompass all major aspects of reading proficiency.

2. ***The measure needs to be reliable and valid.***

Rationale: It is imperative that the reading measures assess reading in a fashion that most professionals consider valid, reliable and comprehensive. In accordance with the recommendations of the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). (1999), these assessments should demonstrate acceptable levels of technical adequacy. At the very least, they should demonstrate reliability of at least .70 and provide evidence of content validity or criterion validity.

Additional criteria for selection of outcome measures for RDD

3. ***Ideally, the measure would include subtests to cover major areas of reading.***

Rationale If so, the study team could analyze the impact on overall reading proficiency, as well as perform exploratory analysis of specific components in reading. Typically, for the primary grades, reading tests include subtests in comprehension, reading vocabulary and aspects of word analysis or word reading.
: Although with RtI, interventions can target any area of reading, our review of the research literature as well as our pilot site visits indicate that many schools target phonemic awareness, phonics/decoding and fluency in their interventions. Thus, it would seem important to explore whether growth is limited to these domains of reading, or whether there is also evidence of growth in comprehension and reading vocabulary. Exploratory analyses of subtests are consistent with this purpose.

4. ***Any standardized, norm referenced measure should have relatively recent norms.***

Rationale: This issue is particularly important for the RDD. In the last 15 years we have witnessed significant effort to improve reading outcomes in the primary grades. In all likelihood, this has resulted in some improvement, which in turn influences national norms. A test normed in 2000, for example, could provide misleading descriptive information. The audience for this study will want to know not only effect sizes for impacts, but also the normative characteristics of students characterized as at risk in the evaluation. Thus, accurate normative information is important.

Reading Achievement Measures for the RDD Design by Grade Level

Grade 1: Our plan is to use the new version of Early Childhood Longitudinal Study-Kindergarten 2011 cohort (ECLS–K) reading assessment. Because ECLS–K will be normed on a nationally representative sample of students who were in kindergarten in spring 2011, normative data will be current allowing study results to placed in a national context. The ECLS-K assessment is computer adaptive test that is individually administered test and designed to take approximately 20-30 minutes. It should provide an accurate gauge of reading performance for 1st graders, especially those who are struggling readers. Our primary focus in this RDD analysis is on students who fall in the at-risk category**.**

Although the technical material on the 2011 ECLS-K is not yet publicly available, we understand from NCES that the new version is designed to address ceiling and floor issues and remains close to the original ECLS-K in design, content coverage and item functioning. The original ECLS-K was based on the NAEP 4th Reading Framework with the addition print and letter awareness. Items for the ECLS-K were taken from widely used existing instruments including the including the Peabody Individual Achievement Test-Revised (PIAT-R), Peabody Picture Vocabulary Test-Revised (PPVT-R), the Primary Test of Cognitive Skills (PTCS), the Test of Early Reading Ability (TERA-2), The Test of Early Mathematics Ability (TEMA-2), and the Woodcock-Johnson Tests of Achievement-Revised (WJ-R). Children respond to between 50 and 70 items that measure the constructs basic skills (print familiarity, letter recognition, beginning sounds, ending sounds, short vowels, long vowels, rhyming words), vocabulary **(**picture-spoken word matching, word recognition) and comprehension (initial understanding, developing interpretation, personal reflection, critical stance). It is a well-designed and psychometrically sound instrument. It directly measures many common emphasis areas in many RtI interventions. We have begun discussions with NCES and the Education Testing Service (ETS), which developed the test, to secure needed permissions for using the test and to train staff in its use.

In addition, we will supplement the ECLS-K with the Test of Word Reading Efficiency-Sight Word Efficiency (TOWRE-SWE), an individually administered test of word reading and fluency. Even though this measure is brief (45 sec), it has strong psychometric characteristics. Moreover, it is simple to administer.

Rationale: Two factors led us to choose a measure of reading efficiency. First, the majority of schools in the study will employ a time-based screening measure such as DIBELS or AimsWEB. Thus, selection of students for tier-2 will be based in part on their ability to perform phonological and non-word reading tasks efficiently, and an outcome measure that reflects fluent reading seems appropriate. Second, many of the interventions for struggling readers focus on building fluent reading. Thus, an adding an outcome measure of efficient word recognition will complement the ECLS-K, which does not address speed of reading.

Grade 2: Our plan is field a test of fluency for second grade, with the TOWRE-SWE as our planned choice. The second grade version of the new ECLS-K will not be available in time for its use in this project. And we are cognizant of the potential burden on study sites if we require too much testing. Hence, our decision to use a short fluency oriented test for second grade, which is closely linked to many of the likely RtI interventions in the study sites and is designed to measure a skill (reading fluency) that is closely linked to reading skills such as comprehension.

Grade 3: The study team proposed to use use state assessment data for each student because of the reduced burden of this option and its policy relevance.

Measures of Student Outcomes for the CITS

For measures of student outcomes for the CITS analysis, the study team will draw on existing data from student records because we need both historical data for the baseline period and follow-up data which also will cover school years prior to 2011-12. Thus, only data from existing student records can feasibility be used for this analysis.

When analyzing impacts on special education identification by disability category and grade retention, we will access student records from participating districts. We expect that district record systems will typically include whether a student has been identified for special education services, but we may have to access specialized records systems to collect information on the disability category of students identified for special education.

For measures of reading achievement, we will access any available student records of reading test scores with pre- and post-RtI start data using a consistent measure. Since testing of all 3rd graders in reading and mathematics was mandated with the 2002 reauthorization of ESEA, 3rd grade state achievement tests are the obvious candidate for this design at that grade level. All states have been administering a 3rd grade reading assessment since at least 2005, allowing for the possibility of a baseline measure of reading performance for at least 3 years prior to RtI implementation. We are also interested in second grade reading test scores but few states have been testing second graders historically. California, Colorado, Florida, and Utah have a history of second grade testing and to the extent our final selection of sites includes enough schools from these states, we will explore the possibility of a second grade CITS analysis for reading.

The study team also plans to analyze impacts on an oral reading fluency measure such as the DIBELS ORF, if it is available historically in enough sites. The r eason s for using the ORF measure as an outcome variable in RDD and CITS analyses outweigh reasons against its use. First, ORF has strong technical properties (i.e., reliability and validity). For example, ORF is strongly correlated with other reading measures (i.e., word identification, text reading, and reading comprehension) and ORF accounts for significant variance on reading comprehension tests (Jenkins, Fuchs, van den Broek, Espin, & Deno, 2002), even after controlling for word identification skill. (See the summary of ORF technical properties below). Second, as a matter of practice, many schools use ORF as an assessment of reading achievement in the primary grades. For example, an estimated 14,000 schools have adopted the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002), which uses ORF as a key reading outcome in Grades 1 and 2 (https://dibels.uoregon.edu/data/index.php) with assessments administered to more than 1,800,000 students (Samuels, 2007). Third, many schools, for better or worse, focus instruction on improving students' oral reading scores, especially those students at risk for poor reading outcomes. This means that ORF should be particularly sensitive to schools' intervention efforts. Fourth, using ORF in RDD and CITS evaluation designs greatly expand the number of schools potentially available for an impact evaluation at Grades 1 and 2. Without ORF, it is unlikely that the study will have any way to estimate impacts at the end of first-grade and few opportunities to estimate impact at the end of second-grade.

The single drawback of using ORF in an impact analysis is the objection by some reading researchers that educators place too much emphasis on this measure, allowing it to warp reading instruction just to raise ORF scores. They regard ORF as a simplistic and narrow measure of reading ability (Goodman, 2006; Valencia, Smith, Reece, Li, Wixson, & Newman, 2010). Thus, it is important to acknowledge that ORF is but one of many valid reading measures and that caution must be exercised in generalizing results based on any single reading measure, including ORF. Further, by incorporating other reading outcomes in impact analyses the study can provide important information about the consistency of results across measures.

B5. Individuals Consulted on Statistical Aspects of Design

Dr. Pei Zhu from MDRC is leading the research design subtask for MDRC. MDRC and SRI have also held multiple conference calls with sub-groups of the study team and with MDRC's technical reviewers for the project, Dr. Howard Bloom and Dr. Marie-Andree Somers. We have also met with our Technical Working Group which has with expertise on RtI designs and analysis on several occasions.

The TWG includes:

- Carol Connor, Florida State University

- Donald Compton, Vanderbilt University
- Judy Elliott, Los Angeles Unified School District
- David Francis, University of Houston
- Paul McDermott, University of Pennsylvania
- Rollanda (Randi) O'Connor, University of California – Riverside
- Amy Sichel – Abington School District (Abington, Pennsylvania)
- Jeff Smith, University of Michigan
- Deborah Speece, University of Maryland – College Park
- Sharon Vaughn, University of Texas – Austin

References:

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91: 444-472.

Angrist, Joshua D., and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *The Quarterly Journal of Economics* 114: 533-575.

Bloom, Howard S. 1995. "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs." *Evaluation Review*, 19(5): 547-556.

Bloom, Howard S. 1999. "Estimating Program Impacts on Student Achievement Using 'Short' Interrupted Time-Series." New York: MDRC.

Bloom, Howard S. 2001. "Measuring the Impacts of Whole-School Reforms: Methodological Lessons from an Evaluation of Accelerated Schools." New York: MDRC.

Bloom, Howard S. 2003. "Using Short Interrupted Time-Series Analysis to Measure the Impacts of Whole School Reform: With Applications to a Study of Accelerated Schools." Evaluation Review 2 1: 3-49.

Bloom, Howard S. 2005. "Randomizing Groups to Evaluate Place-Based Programs." In Howard S. Bloom (ed.). *Learning More From Social Experiments: Evolving Analytic Approaches*. New York: Russell Sage Foundation.

Bloom, Howard S., James Kemple, Beth Gamse, and Robin Jacob. 2005. *Using Regression Discontinuity Analysis to Measure the Impacts of Reading First*. Paper given at the annual American Educational Research Association research conference in Montreal, Canada, April 14.

Bloom, Howard S. 2010. "Modern Regression Discontinuity Analysis." MDRC Working Paper on Research Methodology. New York: MDRC.

Council of Professional Associates on Federal Statistics, *Providing Incentives to Survey Respondents, Final Reports, 1992.* Includes paper by Richard Kulka, at **http://www.copafs.org/reports/providing_incentives_to_survey_respondents.aspx**.

Council of Professional Associates on Federal Statistics, Seminar On Survey Respondent Incentives: Research and Practice, held on March 10, 2008. Including papers by Richard Kulka et.al. and Sandra Berry, et.al. available at http://www.copafs.org/seminars/research_and_practice.aspx.

Deno, S., Marston, D., Mirkin, P., Lowry, L., Sindelar, P., & Jenkins, J. (1982). *The Use of Standard Tasks to Measure Achievement in Reading, Spelling, and Written Expression: A Normative and Developmental Study* (Vol. IRLD-RR-87, pp. 52). Minnesota Univ, Minneapolis Inst for Research on Learning Disabilities.

Dillman, D. A. 1978. *Mail and telephone surveys: The total design method*. New York, NY: John Wiley & Sons.

Fuchs, D., Fuchs, L.S., Mathes, P.G., & Simmons, D.C. (l997). Peer-Assisted Learning Strategies: Making classrooms more responsive to diversity. *American Educational Research Journal, 34*, 174-206.

Goldberger, Arthur S. 1972. *Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations*. Discussion Paper 129-72. Madison: University of Wisconsin, Institute for Research on Poverty.

Groves, R. M., R. B. Cialdini, and M. P. Couper. 1992. Understanding the Decision to Participate in a Survey. *Public Opinion Quarterly, 56*: 475-495.

Hahn, Jinyong, Petra Todd, and Wiler van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression Discontinuity Design." Econometrica 69, 1: 201-209.

James, J. M. and R. Bolstein. 1990. The Effects of Monetary Incentives and Follow-Up Mailings on the Response Rate and Response Quality of Mail Surveys. *The Public Opinion Quarterly*, *54*(3), 346-361.

Jenkins, J. R., Peyton, J. A., Sanders, E. A . & Vadasy, P. F. (2004). Effects of reading decodable texts in supplemental first-grade tutoring. *Scientific Studies of Reading, 8*. 53-85.

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology, 95*, 719-729.

Lee, David S. 2008. “Randomized Experiments from Non-Random Selection in U.S. House Elections.” Journal of Econometrics 142: 675-697.

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*(2), 148-182.

Mathes, P.G., Howard, J.K., Allen, S.H., & Fuchs, D. (l998). Peer-assisted learning strategies for first-grade readers: Responding to the needs of diverse learners. *Reading Research Quarterly, 33*, 62-94

Marston, D. B. (1989). Curriculum-based measurement approach to assessing academic performance: What it is and why do it. In M. R. Shinn (Ed,) *Curriculum-Based Measurement: Assessing Special Children* (pp. 18-78). New York: Guilford.

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., and Long, J. (2009), Curriculum-Based Measurement Oral Reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47*, 427-469.

Schochet, Peter Z. 2008. *Statistical Power for Regression Discontinuity Designs in Education Evaluations*. Washington DC: Institute of Education Sciences, U.S. Department of Education Technical Methods Report (July).

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company.

Thistlethwaite, Donald L., and Donald T. Campbell. 1960. “Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment.” *Journal of Educational Psychology* 51: 309-317 (December).

Vadasy, P., Jenkins, J. R., & Pool, K. (2000). Effects of tutoring in phonological and early reading skills on students at-risk for reading disabilities. *Journal of Learning Disabilities, 33,* 579-590.

Valencia, S., Smith, A., Reece, A., Li, M., Wixon, K, & Newman, H. (2010). Oral Reading Fluency Assessment: Issues of Construct, Criterion, and Consequential Validity. *Reading Research Quarterly*.

van der Klaauw. 1997. *A Regression-Discontinuity Evaluation of the Effect of Financial Aid Offers on College Enrollment*. Working Paper 97-10. New York: C.V. Starr Center for Applied Economics, New York University.

van der Klaauw. 2002. “Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach.” *International Economic Review* 43, 4: 1249-1287 (November).

van der Klaauw. 2008. “Breaking the Link Between Poverty and Low Student Achievement: An Evaluation of Title I?” *Journal of Econometrics* 142: 731-756.