A STUDY OF THE EFFECTS OF USING *CLASSROOM ASSESSMENT* FOR *STUDENT LEARNING* (Study 2.1a)

OMB Clearance Package Supporting Statement

Part B: Collections of Information Employing Statistical Methods

Regional Educational Laboratory
for the
Central Region

Contract #ED-06-co-0023

Submitted to:

Institute of Education Sciences
U.S. Department of Education
555 New Jersey Ave., N.W.
Washington, DC 20208

Submitted by:

REL Central at
Mid-continent Research
for Education and Learning
4601 DTC Blvd., #500
Denver, CO 80237
Phone: 303-337-0990
Fax: 303-337-3005

Project Officer

Sandra Garcia, Ph.D.

Project Director:

Louis F. Cicchinelli, Ph.D.

Deliverable 2007-2.3/2.4

May 31, 2007

This report was prepared for the Institute of Education Sciences under Contract #ED-06-CO-0023 by Regional Educational Laboratory Central Region, administered by Mid-continent Research for Education and Learning. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

TABLE OF CONTENTS

B. COLLECTION OF INFORMATION EMPLOYING STATISTICAL METHODS......................1

1. Respondent Universe and Sampling Methods..1
2. Statistical Methods for Sample Selection...2
3. Methods to Maximize Response Rates and Deal with Non-response...4
4. Test of Procedures or Methods..6
5. Individuals Consulted on Statistical Aspects of the Design..7

references 8

B. COLLECTION OF INFORMATION EMPLOYING STATISTICAL METHODS

1. RESPONDENT UNIVERSE AND SAMPLING METHODS

A purposive sampling strategy will be used to select schools, for several reasons. First, one state was selected in the seven-state region served by the Central Regional Educational Laboratory to allow for the use of data from the statewide achievement tests as the primary outcome; state tests are on different scales and cannot be combined. Two of the Central Region states were ineligible. Nebraska, due to its statewide focus on classroom assessment, and North Dakota, due its statewide fall testing schedule, will be excluded from participation in the study. The state of Colorado was selected as the host state to reflect a long history of standards-based reform and stable state content standards and aligned assessments.

Second, Colorado districts or individual schools that wish to participate in the study need to meet several criteria. Interest in and commitment to forming one or more classroom assessment learning teams with all teachers in Grades 4 and 5 is necessary, for a minimum of three teachers and a maximum of six teachers per learning team. And, student level achievement data linked to teachers will need to be made available. Districts and schools must also understand and agree to the conditions of the study, including random assignment of schools to intervention and control groups.

Based on the minimum eligibility criterion (i.e., enrollment of grades 4 and 5 students), the potential respondent universe includes 945 Colorado schools in 172 districts. About one-third of these schools are located in large or mid-size cities, about one-third are located in an urban fringe area of the cities, and about one-third are located in small towns or rural areas. The schools serve students with different ethnic and socioeconomic backgrounds. Just over half (59%) of the 945 eligible schools are Title I schools. Tables 1 through 4 provide descriptive statistics for the universe of potential schools.

Table 1. Descriptive on the schools in Colorado that enroll grade 4 and 5 students:

N=945	**Total school enrollment (all grades)**	**Total enrollment in Grades 4 and 5**
Mean number of students	393.79	116.79
Standard Deviation	204.39	61.33
Median	387	116

Source: Common Core of Data for the 2004-2005 school year

Table 2. Percent of students eligible for Free or Reduced-Price Lunch (FRL) in Eligible Schools

Percent of total student enrollment eligible for FRL:	Percent of schools:	Number of schools:
Less than 50%	66.7%	630
50% or more	33.3%	315
Total	100%	945

Source: Common Core of Data for the 2004-2005 school year

Table 3. Percent of students of white/Caucasian background in eligible schools:

Percent of total student enrollment that is white:	Percent of schools:	Number of schools:
Less than 60%	36.4%	344
60% to 85%	37.2%	352
85% or more	26.3%	249
Total	100%	945

Source: Common Core of Data for the 2004-2005 school year

Table 4. Eligible schools by student minority enrollment

Percent of total student enrollment that is the indicated racial/ethnic minority:	Percent (Number) of schools:			
	Hispanic	African American	American Indian	Asian American
More than 40%	24.6% (232)	1.5% (14)	0	0
15% to 40%	29.8% (282)	9.6% (91)	0.5% (5)	0.3% (3)
Less than 15%	44.3% (419)	77.7% (734)	83.5% (789)	87.7% (829)
None	1.3% (12)	11.2% (106)	16.0% (151)	12.0% (113)
Total	100% (945)			

School recruitment will involve multiple strategies to develop awareness of and interest in study participation. Networking, broadcasting, and letters of endorsement will be used to establish direct contact with principals and district officials among the eligible schools and districts. Since professional development often involves district support and is included in district-wide capacity building initiatives, we targeted participant recruiting efforts at the district level. A majority of the eligible schools (642) are located in the 20 largest Colorado districts. District recruiting affords a number of advantages, including support of district administration for the study, reducing the number of district-level approvals required, and facilitating access to student-level achievement data.

Ongoing monitoring of characteristics of participating schools will be used to re-direct recruitment efforts to encourage participation among under-represented subgroups of schools; the sample, however, is not a probability sample and therefore is not intended to represent all schools in the respondent universe.

When only one school from a district volunteers to participate in the study, those schools will be placed in a data set. At the time of random assignment, all schools in the data set will be assigned a number via a random number generation procedure such as RNG = MC in SPSS or RANUNI in SAS. Schools in the data set will then be ordered according to the random number. A coin will be flipped (heads = treatment first; tails = control first) to determine with which group to start (treatment or control). Then starting with either treatment or control depending on the results of the coin flip, each school in the data set will be assigned, alternating between treatment and control.

Schools from districts that have more than one school participating will be blocked by district and random assignment will occur within each district using the same method described above. Each school in the district will be assigned a random number. Schools in the district will be ordered by the random number. The first school in the ordered list will be assigned to treatment or control based on the results of a coin toss. The remaining schools on the district list will be assigned, alternating between treatment and control.

Assuming three to six Grade 4 and 5 teachers per school in 64 schools, the sample includes a maximum of 384 teachers and, assuming 60 Grade 4 and 5 students per school, the sample includes a maximum of 3, 840 students. The expected response rates based on standards for longitudinal designs are at least 70 percent for students per classroom/teacher, and at least 90 percent for teachers per school and for schools per group (treatment and control) (National Center for Educational Statistics, 2002).

2. Statistical Methods for Sample Selection

Two power analyses, one on student outcomes and one on teacher outcomes, were conducted in order to determine the sample size necessary to detect the effect of the intent to treat. All power analyses were conducted using Optimal Design software (Liu, Spybrook, Congdon, Martinez, & Raudenbush, 2006), specifically made for power analyses for hierarchical cluster randomized designs. Sample and cluster size were chosen to achieve a high level of power, >0.80. Parameter estimates for the analyses were chosen to be conservative to avoid overestimating power. Rationales for the estimates for effect size, intraclass correlation, and the covariate are described below. Power analyses were conducted for fixed effects.

Random assignment of schools to the treatment or control groups will be blocked by district. However, given the required sample size, the loss of degrees of freedom due to the blocking by district will not have an adverse effect on statistical power. Power analyses were adjusted to reflect the inclusion of one covariate. The anticipate degrees of freedom for the analysis of the main effects is expected to be well above 30 such that the inclusion of strata will have minimal impact on power.

The assumed minimum detectable effect for this study is .25. Given the relatively low financial costs of the intervention—approximately $1500 per school—an effect of this size on student achievement would be worthwhile to detect. An increase of one quarter of a standard deviation in student achievement represents a practically significant effect, equivalent to an increase in 10 percentile points. No empirical evidence is available from field trials of the intervention itself.

Effect sizes from the literature on classroom assessment vary according to the type of assessment intervention and the outcome measure. Black and Wiliam (1998), in their review of the formative assessment literature, report that the effects of classroom assessment on student achievement that typically range from .20 to .30 with some effects as large as .70 or even greater. A recent study on the effects of professional development in classroom assessment found an average effect size of .32 after six months of teacher training (Wiliam, Lee, Harrison, & Black, 2004). A study of 5th and 6th grade math students found an effect of .40 for effects of student self-evaluation, which is included in *CASL* via student involvement (Ross, Hogaboam-Gray, & Rolheiser, 2002).

A conservative value of 0.15 was selected for the intra-class coefficient (ICC) based on the following sources. Raudenbush et al. (2006) cite typical intra-class correlation coefficients for educational achievement to be between 0.05 and 0.15. Schochet (2005) states that ICC for standardized test scores often range between .10 and .20. Bloom et al. (2005) found ICCs in Grade 5 reading and math ranged from .12 to .29 across five different districts.

Prior achievement was selected as a cluster-level covariate, and the proportion of post-intervention variance explained by pre-intervention test scores of .50 was deemed an appropriately conservative estimate given the correlations between prior and future achievement based on findings in the research literature. Schochet (2005) concludes that the proportion of variance explained by pretest measures is at least .50 when student level data are used. In a 1999 study Bloom et al. (1999) found similar values. In a later study Bloom et al. (2005) found values ranging from .33 to .81 across five districts for school level pretests.

A power analysis for the outcome of student achievement was conducted using the above parameter values as well as the following very conservative estimated sample sizes at the end of the study: 60 students were assumed to be nested within each school. This accounts for student mobility and the potential attrition of teachers within schools; a sample size of 60 students assumes 15 students per classroom and four classrooms per school. Given the above assumptions, Optimal Design software calculated that 47 clusters (approximately 24 intervention and 24 control clusters) were necessary to achieve the desired power of >.80 for the student achievement outcomes.

Power analyses were also conducted in order to determine the sample size necessary to detect the intention to treat on teacher outcomes. Parameter estimates for effect size, intraclass correlation, and proportion of post-test variance explained by the baseline measure were chosen for the following reasons.

Few rigorous studies have explicitly examined the effects of professional development on teachers. A study of teacher assessment competencies using a national sample found an effect size of .20 on a test of knowledge favoring teachers who had taken a graduate level measurement course over teachers who had not taken a course (Plake, Impara, & Fager, 1993). A study using the same instrument found that teachers' scores increase an average of one standard deviation after taking a graduate course in assessment (O'Sullivan & Johnson, 1993). This same study found a difference of two standard deviations when comparing scores on classroom assessment performance tasks between teachers who had completed a graduate course in assessment and teachers who had not completed a course. Evaluations of the effects of training in standards-aligned classrooms found effects ranging from approximately .50 to 1.00 for the effect on teachers' familiarity and use of standards in instruction and assessment (Wolfe & Jarvinen, 2002, 2003). An estimated effect size of .50 was assumed based on the above findings.

Little empirical evidence could be found regarding estimates of intraclass correlations or covariation for teachers. A value of .10 was used as the estimate of the ICC based on the assumption that there would be slightly less shared variance between teachers than between students. A conservative value of .20 was assumed for the correlation between teachers' baseline scores on the test of assessment knowledge and teacher outcomes for several reasons. First, we assumed teachers' scores would likely be relatively unstable due to variations in implementation of the training as well as variations in other professional development across the schools. Second, the baseline scores on the test of assessment knowledge will be used as the covariate for all teacher outcomes and the correlation between this measure and the other outcome measures in not known.

In addition to the above, an assumption of four teachers per cluster was used to estimate final sample size for the power analysis for teacher effects. This analysis was also based on an estimated effect size of .50, a proportion of post-intervention variance explained by pre-intervention test scores of .2, and an intra-class correlation of .10. Using the above assumptions, Optimal Design calculated that 41 clusters were necessary to achieve a power of >.80.

The target sample size was determined by the need for sufficient power to detect the intervention effect on student achievement and to account for attrition. A sample of 24 intervention and 24 control schools is needed to achieve the needed statistical precision of the impact estimates for both students and teachers. Assuming approximately 25% attrition we will need 32 clusters for the intervention and control groups, making a total of 64 schools in the target sample.

For information regarding whom the sample represents, please see Response B1. Information about the instruments themselves and about the data collection schedule is in Response A2. A description of the quality control procedures is in Response B4.

3. Methods to Maximize Response Rates and Deal with Non-response

The primary outcome for this study is student achievement; student scores from the state-wide achievement test administered under NCLB will be used as the data source for this outcome. A response rate above 95% is expected because NCLB requires all students to participate in the state assessment system. Teacher response rates are anticipated to be above 80% based on experience and similar research studies. For example, in a study of professional development in classroom assessment achieved 79% of teachers initially enrolled in the study responded to requests for data (Wiliam et al., 2004). In a different study of professional development in vocabulary instruction, a response rate of 93% was achieved among teachers, all of whom had volunteered to be part of the study (Apthorp, 2006). All the teachers in this study will be volunteers and will have agreed to the requirements for data collection prior to beginning the study.

Several steps will be taken to help maximize response rates. First, data collection instruments will be administered online as much as possible. The online nature of the data collection will facilitate data collection by eliminating the need to deal with paper documents or mailing activities. Second, data collection using online instruments will be managed electronically. Reminders about upcoming and current data collection activities will regularly be sent to participating teachers via email. Two weeks before each data collection is due, teachers will receive an email message providing them with a link to the instrument and a requested timeline for completion. The full data collection schedule will be communicated to respondents at the onset of the study. The advance schedule, reminder, and response window structure will allow participants to plan and to incorporate the data collections into their schedules.

Because of the numerous stressors on teachers' time and the importance of retaining participants throughout the course of the study, we believe it is imperative to compensate participants for the burden of response to improve response rates and maintain data quality (Office of Management and Budget, 2006a, 2006b). Participating teachers will receive the following:

- Acknowledgment of each participant's selection as a professional honor and an opportunity to contribute to knowledge in the field with a certificate of award at the onset of the study.
- At the conclusion of the study, acknowledgment of each participant's professional contribution to the knowledge in the field with a certificate of award signed by researchers and a recognized national expert in classroom assessment.
- Participant briefing booklet on findings of the study.
- Opportunity to be awarded an expense-paid trip to Assessment Training Institute 2009 or other comparable summer conference on classroom assessment, with one such award made by lottery to each group of study participants. Winning tickets will be drawn from a pool; one ticket for entry to the pool is earned for each on-time response to data collection requests.
- Compensation for response burden, payable according the time required. Compensation will be spaced to correspond with the four main data collection periods of the study described above. Total compensation for data collection will be $300 per teacher participant to cover the anticipated minimum of eight hours of response burden. Four payments of increasingly larger increments will be made to encourage participants to provide data throughout the entire study.
- Satisfaction of participation in a project with high visibility.
- Potential benefit of intervention to participant's students (intervention schools initially, all groups eventually).
- Potential for satisfaction from professional growth.
- Potential for satisfaction from working with other teachers interested in classroom assessment.
- Visibility with and support of participant's principal, whose signature of support is part of the school application for the study.

We plan to deal with non-response via several methods. First, teachers who do not respond to initial requests for responding will receive follow-up reminder emails to encourage their completion of instruments. Second, a sufficient sample of schools will be recruited to provide reliable data in the event of respondent attrition. Power analyses indicate that a sample of 47 schools, approximately 24 intervention and 24 control, is adequate to achieve the desired level of power in the analysis of changes in student achievement (see question B2 for more detail). Given this requirement, 64 schools will be recruited (32 intervention and 32 control), allowing for an attrition rate of up to 25%. This study uses a purposive sample; no plans are in place to weight the sample to represent subgroups.

4. Test of Procedures or Methods

As much as possible, existing instruments with documented reliability and validity were identified for use in this study. Existing measures were selected with construct validity in mind. Existing instruments were adapted as necessary to help ensure sensitivity to the intervention. When existing measures were not available, original instruments were developed specifically to provide the necessary alignment with and sensitivity to the intervention.

All instruments—original and existing—were reviewed and are being pilot-tested. Instrument review was used to help ensure that the instruments measure constructs representing agreed-upon definitions in the research literature and will be sensitive to intervention effects. Pilot testing with a small sample (under 10) of individuals from the target population is being conducted to help to ensure that directions are clear, requests for data are unambiguous, the process is efficient, the time required is known in advance, and online instruments function properly and are easy to use. A report on the pilot test is not yet available.

The first opportunity to collect data from a large sample of respondents will be in Fall 2007. For this reason, data collected in Fall 2007 will be used to examine the psychometric properties of the test of assessment knowledge, the teacher work sample, and the survey of student involvement. The data will be used to conduct a number of analyses such as factor analysis, item statistics, and composite score statistics (e.g., item score frequency distributions, item means and standard deviations, item-total correlations, composite score frequency distributions, composite score means and standard deviations, composite score intercorrelations, and internal consistency).

For the test of teacher knowledge of classroom assessment, approximately 50 items will be administered at baseline. Items will be considered for dropping if they are not functioning well or do not contribute to the reliability and/or validity of the composite score. Even if one-third of the 50 items administered are dropped, 33 items will remain. Scores from a 33-item test will very likely provide adequate reliability and validity for the intended purpose.

In addition to the test of assessment knowledge, the teacher survey of student involvement also will be administered in the fall of 2007. Data from these surveys will be used to examine the psychometric properties of the survey and its items. Fall 2007 data from the teacher work samples will be used to calculate inter-rater agreement for the ratings of teacher work samples.

5. Individuals Consulted on Statistical Aspects of the Design

The statistical aspects of the design have been reviewed thoroughly by staff at the Institute of Education Sciences, as well as by members of the study’s expert panel listed in Section A.8. The following individuals have worked closely in developing the statistical procedures and will be responsible for data collection and data analysis:

Dr. Bruce Randel, Principal Investigator, 303-632-5576

Dr. Helen Apthorp, Co-Principal Investigator, 303-632-5622

Dr. Andrea Beesley, Study Director, 303-632-5541

REFERENCES

Apthorp, H. A. (2006). Effects of a supplemental vocabulary program in third-grade reading/language arts. *The Journal of Educational Psychology, 100*(2), 67-79.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, October*, 139-148.

Bloom, H. S., Bos, J. M., & Lee, S.-W. (1999). Using cluster random assignment to measure program impacts: Statistical implications for the evaluation of education programs. *Evaluation Review, 23*(4), 445-489.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2005). *Using covariates to improve precision: Empirical guidance for studies that randomize schools to measure the impacts of educational interventions. MDRC Working Papers on Research Methodology*. New York: Manpower Demonstration Research Corp.

Liu, X., Spybrook, J., Congdon, R., Martinez, A., & Raudenbush, S. (2006). Optimal Design for Multi-Level and Longitudinal Research (Version 1.77): HLM Software.

National Center for Educational Statistics. (2002). NCES Statistical Standards (NCES 2003-601). Retrieved March 19, 2007, from http://nces.ed.gov/pubs2003/2003601.pdf

O'Sullivan, R. G., & Johnson, R. J. (1993). *Using performance assessment to measure teachers' competence in classroom assessment.* Paper presented at the Annual meeting of the American Education Research Association, Atlanta, GA.

Office of Management and Budget. (2006a). *Questions and answers when designing surveys for information collection*. Washington, D.C.: Author.

Office of Management and Budget. (2006b). *Standards and guidelines for statistical surveys*. Washington, D.C.: Author.

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice, 12*(4), 10-12, 39.

Raudenbush, S., Spybrook, J., Liu, X., & Congdon, R. (2006). Optimal design for longitudinal and multilevel research: Documentation for the "Optimal Design" software. Retrieved June 1, 2006, from http://sitemaker.umich.edu/group-based/files/odmanual-20060517-v156.pdf

Ross, J. A., Hogaboam-Gray, A., & Rolheiser, C. (2002). Student self-evaluation in grade 5-6 mathematics effects on problem-solving achievement. *Educational Assessment, 8*(1), 43-59.

Schochet, P. Z. (2005). *Statistical power for random assignment evaluations of education programs*. Princeton, NJ: Mathematica Policy Research.

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education Principles Policy and Practice, 11*(1), 49-65.

Wolfe, E., & Jarvinen, D. (2002). *Standards-aligned classroom initiative: Year 2 evaluation report*. Springfield: Illinois State Board of Education.

Wolfe, E., & Jarvinen, D. (2003). *Standards-aligned classroom initiative: Year 3 evaluation report*. Springfield: Illinois State Board of Education.