The Council of Chief State School Officers

EFFECTS OF TEACHER PROFESSIONAL DEVELOPMENT ON GAINS IN STUDENT ACHIEVEMENT

How Meta Analysis Provides Scientific Evidence Useful to Education Leaders

LEADERSHIP, ADVOCACY, SERVICE
CCSSO
THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

June 2009

THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

State Education Indicators

The Council is a strong advocate for improving the quality and comparability of assessments and data systems to produce accurate indicators of the progress of our elementary and secondary schools. The CCSSO education indicators project is providing leadership in developing a system of state-by-state indicators of the condition of K–12 education. Indicators activities include collecting and reporting statistical indicators by state, tracking state policy changes, assisting with accountability systems, and conducting analyses of trends in education.

The meta analysis study is supported by a grant from the National Science Foundation, Division of Research on Learning in Formal and Informal Settings (Award No. #REC-0635409). A draft of this paper was first presented at the Annual Meeting of the Educational Research Association, Division K Teaching and Teacher Education in April 2009 in San Diego, California.

Rolf K. Blank, Director of State Education Indicators

Paper copies of this report may be order for $10 per copy from:

Council of Chief State School Officers
One Massachusetts Avenue, NW, Suite 700
Washington, DC 20001-1431
Phone (202) 336-7000
Fax (202) 408-8072
www.ccsso.org

ISBN: 1-933757-09-4

Effects of Teacher Professional Development on Gains in Student Achievement

How Meta Analysis Provides Scientific Evidence Useful to Education Leaders

Rolf K. Blank
Nina de las Alas

June 2009

Report prepared under a grant to the Council of Chief State School Officers from the National Science Foundation, Grant #REC-0635409

Project title : "Meta Analysis Study of the Effects of Teacher Professional Development with A Math or Science Content Focus on Improving Teaching and Learning"

Council of Chief State School Officers, Washington, D.C.

Abstract

The Council of Chief State School Officers (CCSSO) was awarded a grant from the National Science Foundation to conduct a meta analysis study with the goal of providing state and local education leaders with scientifically-based evidence regarding the effects of teacher professional development on improving student learning. The analysis focused on completed studies of effects of professional development for K-12 teachers of science and mathematics. The meta analysis results show important cross-study evidence that teacher professional development in mathematics does have significant positive effects on student achievement. The analysis results also confirm the positive relationship to student outcomes of key characteristics of design of professional development programs.

Acknowledgements

Several individuals helped made this study and report possible. We would like to recognize Andrew Porter, Dean of the Graduate School of Education, University of Pennsylvania, and Betsy Becker, Professor in the College of Education, Florida State University, for their invaluable advice and assistance in marshalling the development of the study design and the ensuing data analyses. We would like to thank Kwang Suk Yoon of the American Institutes for Research (AIR) for his technical assistance and advice with the study coding form and process, and we greatly appreciate his strategic assistance in leading the training for coders. We appreciate the statistical analysis assistance of Ariel Aloe (Florida State University) and Michelle Peters (George Washington University). We also wish to express appreciation to the team of graduate students who diligently coded nearly 75 reports eligible for inclusion in the meta analysis: Katie Canatsey, Ryan Fink, Rae Seon "Sunny" Kim, Kavita Mittapalli, Michelle Peters and Breena Welker. Finally, we acknowledge the support of the National Science Foundation's, Division of Research, Evaluation and Communication in the Education and Human Resources Directorate, in awarding the study grant to the Council of Chief State School Officers.

The views and opinions expressed in this report and any errors in judgment or fact rests solely with the authors.

Table of Contents

Effects of Teacher Professional Development on Gains in Student Achievement: How Meta Analysis Provides Scientific Evidence Useful to Education Leaders

In the present education policy environment a high priority has been placed on improving teacher quality and teaching effectiveness in U.S. schools (Darling-Hammond et al., 2009; Obama, 2009). Standards-based educational improvement requires teachers to have deep knowledge of their subject and the pedagogy that is most effective for teaching the subject. States and school districts are charged with establishing and leading professional development programs, some with federal funding support, which will address major needs for improved preparation of teachers. The whole issue of teacher quality, including teacher preparation and ongoing professional development, and improving teacher effectiveness in classrooms, is at the heart of efforts to improve the quality and performance of our public schools.

The Council of Chief State School Officers (CCSSO) has led recent initiatives designed to identify, analyze and disseminate important findings from research and evaluation studies of teacher professional development. Our goal is help K-12 education decision-makers base their decisions on programs using best evidence of effectiveness (Blank, et al, 2007; 2008; http://www.ccsso.org/projects/improving_evaluation_of_professional_development). In 2006, CCSSO was awarded a grant from the National Science Foundation (NSF) to conduct a meta analysis study with the goal of providing state and local education leaders with scientifically-based evidence regarding the effects of teacher professional development on improving student learning. The analysis has focused on completed studies of effects of professional development for teachers of science and mathematics. The two-year study was designed to measure and summarize consistent, systematic findings across multiple studies that show significant effects of teacher professional development on student achievement gains in K-12 mathematics or science. The present paper provides a summary of findings from the CCSSO meta analysis. In the paper we describe the studies that met the criteria for inclusion in the meta analysis, and explain the steps in the meta analysis methodology as applied in this area of education research. Meta analysis is frequently used in fields such as medicine, mental health, and criminal justice to confirm and validate findings across studies. Our paper helps to demonstrate why effect sizes and meta analysis are important for comparison of findings across education research and evaluation studies to adequately determine the quality and significance of evidence concerning a key education policy issue, such as designing and implementing teacher professional development.

State Education Leader Needs for Research Evidence

State education agencies are responsible for directing and managing the use of federal funds for teacher development and improvement as well as guiding programs supported by states. Additionally, states are now required under NCLB to report on the qualifications of teachers in core academic subjects and the proportion of teachers that receive high quality professional development each year. Finally, state education agencies provide leadership for local systems on how to design, select, and implement professional development for teachers. Strong, research-based program designs, and evidence on their effects, are now in high demand across the U.S.

State responsibilities for administering, designing, evaluating, and reporting on federally supported and state-funded programs for improving teaching and teacher quality provide a strong rationale for the proposed work by CCSSO to lead a meta analysis study of effects of well-designed professional development programs. States and in turn local districts seek models for designing and implementing effective professional development and particularly models supported by research evidence.

The CCSSO meta analysis study of effects of professional development with mathematics and science teachers is important for state education leaders because of four intersecting trends that are now strongly affecting education policy, data, and research.

1) **Federal legislation**. NCLB pushes for use of scientifically-based research in program decisions and evaluation of effectiveness of programs.

2) **Student achievement as the preferred measure of effects of programs**. The increasing interest within the education community and from policymakers for measuring effectiveness of initiatives by evidence of gains in student achievement, partly due to the improved capacity of data systems to relate programs to student outcomes.

3) **Recent research findings**. A large body of research has identified the design and features of professional development for teachers which will be more likely to produce effects on student learning.

4) **State leadership needed with teacher development resources**. Typically, we see a small state policy role in the design and evaluation of professional development, and local program designs are not often based on research evidence and thus may be lacking coherent or consistent focus.

Federal legislation supporting funding for K-12 public education under No Child Left Behind (NCLB) has produced a strong push toward application of results from scientifically based research in education program decisions and methods of evaluation. NCLB regulations call for programs that have been proven effective through scientifically-based research (Shavelson & Towne, 2002). In implementing NCLB through the several Title programs, the U.S. Department of Education has advocated for program evaluations that are based on experimental designs. A challenge for state education agencies has been to carry out their legislated function of directing federally funded programs for teacher improvement that meet criteria for quality as specified under NCLB Title II (Birman, et al, 2007). States have also been challenged in determining how to encourage and fund evaluation studies that use experimental designs, especially those with random control trials, and would meet the goal of providing scientific evidence of the effects of teacher-focused improvement efforts on improving the achievement of students they teach (Noyce, 2006; Coalition for Evidence-Based Policy, 2003).

Under the Title IIB Math-Science Partnership program of NCLB, program grants are awarded by state competitions. State education agencies are responsible for ensuring that programs include scientifically-based evaluations of program outcomes as well as reporting program results to the U.S. Department of Education. Reviews of existing program evaluations indicate that most professional development programs for math and science teachers are not evaluated with experimental designs (CCSSO, 2006; Frechtling, 2001). States and districts currently have very

limited capacity for relating pre-service teacher preparation or professional development to student outcomes (Carey, 2004).

Student achievement as the preferred outcome measure. Education research that measures effects of improving teacher preparation and development of teacher knowledge and skills on change in student achievement has developed and expanded since the 1990s. Kennedy carried out one of the first reviews of research on the relationship of quality of teacher preparation to subsequent student achievement a decade ago (1998). At that time, she identified a relatively small number of research studies that were able to draw a direct link between the level of teacher preparation in their teaching field and achievement of students. Darling-Hammond (1999) analyzed large-scale assessment data across the states, and her research results showed that teacher preparation in field was positively related to student achievement. These study findings resulted in extensive policy and research debate, that still continues, about the importance of formal teacher preparation and qualifications, including teacher certification.

More recently, several major research synthesis projects have broadly analyzed evidence on the effects of mathematics and science teacher preparation and development initiatives on student achievement. One approach to reviewing evidence across studies is to apply a logic model and to examine the relationship of teacher preparation on student achievement through effects on intervening variables such as teacher knowledge and instructional practices (Clewell et al., 2004; Ingvarson, Meiers & Beavis, 2005). This kind of full analytic model allows educators and leaders to identify key decisions about the organization, delivery and support of teacher development that are ingredients to positive outcomes.

Another approach to research synthesis analysis is to specifically define teacher professional development initiatives and to identify those studies which reveal effects on student achievement directly linked to the initiative. In research for the Southwest Regional Education Lab, Yoon and colleagues (2007) reviewed findings from several thousand studies on the effects of teacher professional development programs and initiatives to determine evidence of effects on student achievement. The synthesis identified relevant findings by applying the ED/IES What Works Clearinghouse criteria for experimental design and measuring effect size. This synthesis identified nine studies that met the criteria in the published research literature, and all nine studies were based on small numbers of teachers and measurement of change with achievement tests closely aligned to the treatment model. A new paper by Wayne, Yoon and AIR colleagues (2008) describes in detail how experimental designs can be used to analyze outcomes from teacher preparation and development.

Recent research on effective teacher development. A large body of education research has been published over the past decade which provides a base of knowledge about the characteristics of effective programs of teacher professional development in mathematics and science. The rationale for recent federal policy toward teacher professional development through NCLB and through NSF education programs has cited findings from research documenting characteristics of initiatives for teacher development that were proven effective in improving teaching (Garet et al., 1999; Hiebert, 1999; Loucks-Horsley et al., 1998; Corcoran & Foley, 2003; National Commission on Teaching & America's Future, 1996; Weiss, et al., 2001; Guskey, 2003; Showers, Joyce, & Bennett, 1987; Supovitz, 2003). There is also extensive published research focusing on the role of teacher knowledge and skills in student learning, the kinds of knowledge teachers need, and what knowledge is critical to effective teaching (e.g., Ball

& Bass, 2000; Borko, 2004; Hill, Schilling & Ball, 2004; Wilson & Berne, 1999; Hill, Schilling & Ball, 2004; Ball & Bass, 2000).

Although there has been strong research evidence that could contribute to improving teacher professional development methods and delivery, there still exists a significant gap in translating research into practice. Results from large-scale national studies early in this decade indicate that most professional development initiatives for teachers are not designed to meet the key characteristics of effectiveness we now recognize from research (Corcoran & Foley, 2003; Garet et al., 2001; Desimone et al., 2002; Corcoran & Foley, 2003Garet et al., 2001).

Improve state leadership. The current state role in setting policies and providing leadership for high quality professional development is weak in many states—that is, states may provide broad guidance but leave the definition, design and delivery of programs and teacher development services to districts, regional service agencies, or other providers (Corcoran, 2007). Currently most program decisions are left to school leaders or to individual teachers regarding types of professional development, course credits for re-licensure, or pay and promotion. The existence of different levels of responsibility for professional development and multiple sources of funding have produced a fragmented, non-targeted system of development of teachers (Birman & Porter, 2002; Choy et al., 2006; Correnti, 2007; Choy et al., 2006; Hezel Associates, LLC, 2007; Birman & Porter, 2002).

Miles, Odden, Fenmanich, and Archibald (2004) studied the total costs of professional development across a large sample of districts and found that an average of $4,380 is spent annually per teacher. Case studies of six districts indicate mixed results from investments in professional development (Chambers et al., 2008). Education systems are allocating extensive funds to professional development. While most teachers do receive some professional development each year, measurable effects are hard to demonstrate due to lack of consistency, content focus and coherence among the professional development activities provided.

States can improve the use of resources and increase their policy role with teacher professional development through reference to findings from research and evaluation. Research on the state role in teacher development has been mostly limited to case studies of specific state initiatives or policies, and organizational characteristics related to program delivery (e.g., Cohen & Hill, 2001). Teacher education and professional development programs conducted by institutions or providers supported by states and districts are evaluated as separate entities, and evaluation criteria and methods are diverse. Policymakers thus find it difficult to gain a comprehensive picture of what works best in improving teacher skills and knowledge or even what effect different amounts of coursework or pre-service education make a difference in improving teaching.

Also, states now have better access to data for measuring effects of programs on student achievement. NCLB did and still continues to provide funding and support for statewide integrated data systems with student and teacher records that provide for longitudinal analysis of student achievement and measuring improvement from grade to grade, and about half the states have received competitive grants to improve longitudinal data systems (National Center for Education Statistics, n.d.). States and districts are in a better position to employ large data bases to analyze the effects of specific program interventions, such as teacher professional development, than they were even three years ago. Now, analysis of state data from education

information systems is supported by a new federally-supported center—National Center for Analysis of Longitudinal Data in Education Research or CALDER (Harris & Sass, 2007).

Study Questions

The CCSSO meta analysis study focused on identifying research from recent studies that measure effects of teacher professional development with a content focus on math or science. The meta analysis was carried out to address two primary questions:

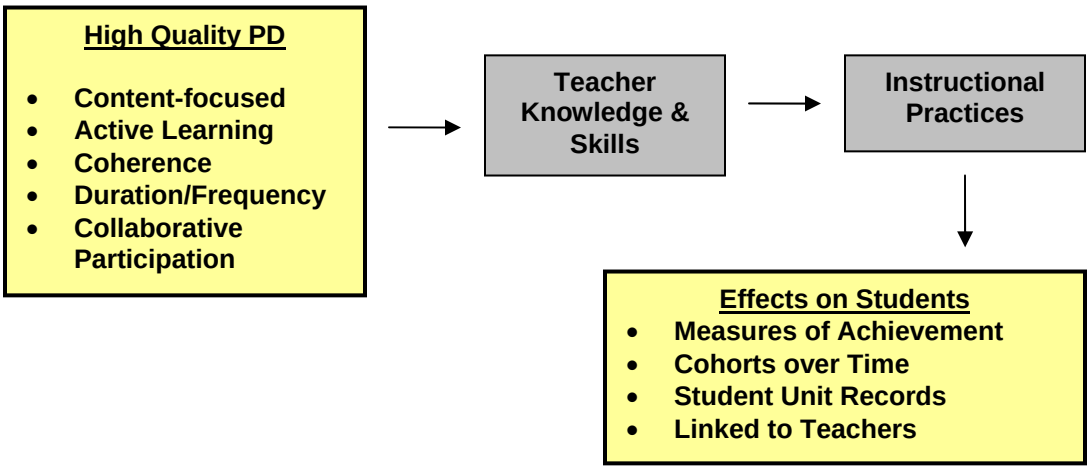
1) What are the effects of content-focused professional development for math and science teachers on improving student achievement as demonstrated across a range of studies?

2) What characteristics of professional development programs (e.g., content focus, duration, coherence, active learning, and collective participation of teachers) explain the degree of effectiveness, and are the findings consistent with prior research on effective professional development (e.g., content focus, duration, coherence, active learning, and collective participation of teachers)?

One goal of the present paper is to report on the results of the meta analysis which has been completed by the CCSSO study team. A second goal of the paper is to report on the use of meta analysis as a method for providing evidence for education decision-makers. The paper describes the methodology developed and carried out by the CCSSO team. With the current needs of education leaders for scientifically-based research evidence of program effects, we can report on the process for conducting this meta analysis as an important outcome of the study. The study results also include a set of common criteria for identifying studies demonstrating significant effects and how statistical procedures are used to establish acceptable effect sizes across a range of studies with varying treatments, sample sizes, and outcome measures. The paper will outline the meta analysis steps toward identifying accepted studies and effects, and then describe the important programmatic findings gained from the studies.

The meta analysis study data collection follows the broad logic model for evaluating professional development developed in previous CCSSO projects (see **Figure 1**). In particular, the meta analysis study design centered on two areas: capturing the characteristics of the professional development models discussed in the studies, and documenting the resulting measurable student outcomes the studies attribute to the professional development programs.

Figure 1: Logic Model



Study Design

The design for the CCSSO meta analysis built on prior studies in education (Borman et al., 2002; Yoon et al., 2007; Lipsey & Wilson, 2001) and applied it to findings about professional development across states and districts. The study design had four basic steps:

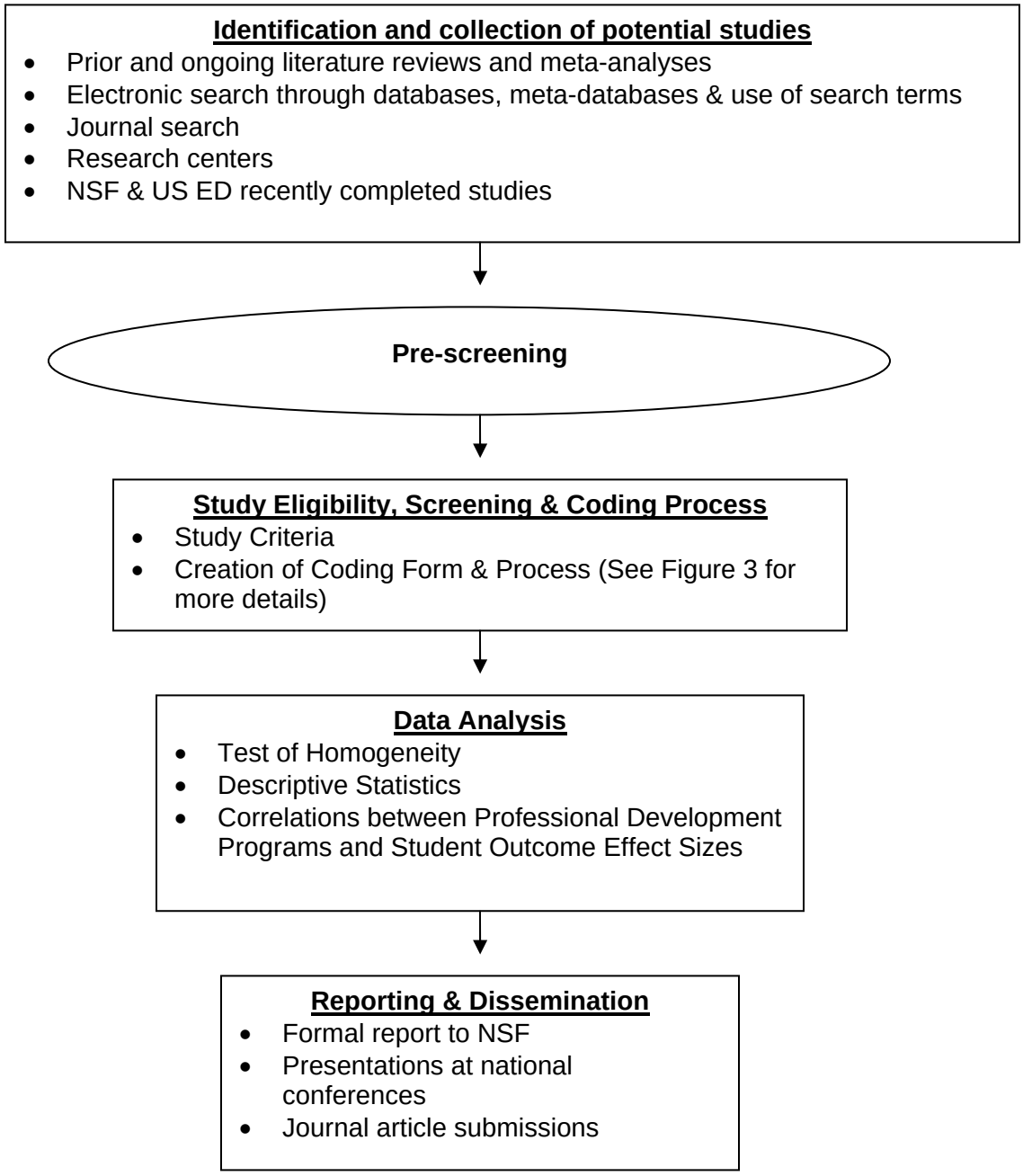
1) identification and collection of potential studies,
2) determination of study eligibility and conduct coding process,
3) data analysis, and
4) reporting and dissemination.

Figure 2 illustrates the process in more detail.

At the start of the CCSSO meta analysis, discussions with researchers from the American Institutes for Research (AIR) who were conducting a teacher professional development systematic review for the Southwest Regional Educational Laboratory (REL Southwest) precipitated adjustments in the literature search for the CCSSO study. Whereas the AIR-REL Southwest project focused on only published studies that cover reading/English language arts, mathematics and science from Australia, Canada, the United Kingdom and the United States, we widened our literature search to include unpublished works and yearly evaluation reports from ongoing projects.

From May through November 2007, we conducted an intensive electronic search using the following databases and meta-databases: ERIC, PsycINFO, ProQuest, EBSCO host Academic Premier Search and Education Abstracts, Dissertation Abstracts, and the database of the Campbell Collaboration. Search words used included "professional development," "staff development," "math," "science," "research study," and "student achievement." We also reviewed the online database Teacher Qualifications and Quality of Teaching (http://ott.educ.msu.edu/tqqt/).

Figure 2: Overview of the study design



In addition, searches were conducted targeting certain periodicals, namely, *Review of Educational Research, Educational Evaluation and Policy Analysis, Education Policy Analysis Archives, TC Record, Journal of Research in Science Teaching, Science Education, Electronic Journal of Science Education, Research in Science & Technological Education, Journal of Science Education and Technology, Electronic Journal of Literacy Through Science, Taylor and Francis Group of scholarly periodicals, Journal of Chemical Education, ERS Spectrum, and School Science and Mathematics.* Journals from associations such as the National Association for Research in Science Teaching, the Association for Science Education, and the American

Educational Research Association (AERA) were reviewed. With the latter, additional searches were conducted among the 2007 annual meeting abstracts to identify potential documents. CCSSO also examined the publications and databases of key research centers including RAND, Research for Better Schools, the Center for Research on the Education of Students Placed At Risk (CRESPAR), Consortium of Policy Research in Education (CPRE) and the Center for Comprehensive School Reform and Improvement. Lastly, CCSSO solicited principal investigators listed in USDOE Teacher Preparation Continuum, NSF MSP projects, IES funded projects, and Local Systemic Initiative (LSI) study sites.

Cross-checks were also conducted using the prior reviews in teacher professional development. In particular, documents that were identified in the AIR-REL Southwest studies that passed its prescreening phase, reports that were included in the review conducted by Abt Associates (O'Reilly & Weiss, 2006; Scher & O'Reilly, 2007) and in the seminal Kennedy review (1998) were highlighted for inclusion.

As a result, 416 reports were identified for pre-screening. A review of their abstracts eliminated 82 percent or 342 reports because they were deemed irrelevant based on the pre-screening criteria (see **Table 1**). The remaining 74 documents were screened by a team of trained coders.

Table 1: Pre-Screening Criteria

Criterion	Description
Topic Focus	The document discusses the effects of inservice teacher professional development on student learning.
Population Focus	The study sample focused on teachers of mathematics and/or science and their students in grades K-12.
Study Design	The document discusses an empirical study.
Outcomes	The document must report direct student achievement outcomes, not distal student outcomes such as feelings, impressions or opinions from students about their learning.
Time Frame	The document had to be released between Jan. 1, 1986 and August 31, 2007.
Country	The study had to take place in the United States.

Coding Form

We adopted the coding form and reconciliation form used by AIR in their review (see **Appendix A**). The coding form was a systematic template that simultaneously assisted coders in classifying the pool of potential studies for inclusion as well as collect information from each study that was used in the meta analysis. The coding form appeared as an Excel file with multiple spreadsheets. A coder used the first spreadsheet to record his or her determination that the document 1) presents an empirical study with quantitative data on an in-service professional development program for teachers of math and/or science and includes student achievement outcomes; 2) uses a research design that produces valid and measurable results; 3) reports at least one effect size or provides sufficient data to compute at least one effect size; and 4) records some professional development characteristics. At each step (see **Figure 3**) the study was sorted for or against further consideration and inclusion. Subsequent spreadsheets in the file collect data that were used for the meta analysis: student and teacher outcome measures, sample sizes of teacher

and student populations, teacher and student characteristics, attributes of the professional development program or initiative, and statistical data from the study's results that when entered will automatically compute effect sizes based on the student outcome measures. Information on the completed coding sheet was transferred to the reconciliation form which recorded input from both coders assigned to review the document. A third member reconciled any conflicting codes recorded by the coders and presented the final data through the reconciliation form to be entered for data processing and analysis.

We recruited and trained a cadre of graduate students (mostly doctoral students in education and statistics) from Florida State University, George Mason University and George Washington University to code the 74 pre-screened documents. Initial full-day training was followed up a week later with a one-hour post-training session to gauge coders' comfort level with the task on-hand and addressed any lingering general questions about the coding process. At the end of the full-day training session, coders were assigned to specific documents and to work in rotating teams of two. Coding and reconciling assignments are rotated throughout the coding process to maintain independent and unbiased reviews. Using a password-protected open source online portal run by Liferay called "Communities," the coders worked remotely and asynchronously and posted their results onto a common area. Disputes in coding were settled by having an assigned reconciler who made final judgments for each question in the coding form. Questions and comments during the process of coding were conducting either over the phone, email, or in the communities comment page.

Figure 3 illustrates the three stages of coding each pre-screened document underwent and the resulting documents that cleared each step.

Figure 3: Flow of Documents Reviewed and Included in the Meta Analysis Study

Stage I, Pt. 1	*To determine if a document meets the following criteria: empirical quantitative research paper on an in-service PD program for teachers of math and/or science with student achievement outcomes*
Total # of Documents	74
# of Documents Rejected	41
# of Documents Passed	**33**
Inter-rater Reliability Rate (=60/74)	**0.81**

Stage I, Pt. 2	*To determine if a document's research design would validly produce measurable results*
Total # of Documents	33
# of Documents Rejected	11
# of Documents Passed	**22**
Inter-rater Reliability Rate (=30/33)	**0.91**

Stage II	*To determine if a document has enough data to compute an effect size or reports an effect size*
Total # of Documents	22
# of Documents Rejected	2
# of Documents Passed	**20**
Inter-rater Reliability Rate (=21/22)	**0.95**

Post-coding	*To determine comparable effect sizes for each document for meta analysis*
Total # of Documents	20
# of Documents Rejected*	4
# of Documents Passed	**16**

The flow chart shows that 55 percent of the 74 documents that passed the pre-screening criteria failed at Stage I, Part 1, primarily because the documents did not meet the criteria. For example, one document that did not continue to the next round of screening focused more on a comparison of two curricula programs and not on professional development.

A third of the documents failed to move on to Stage II, mainly for shortcomings in meeting the criteria for any of the four types of eligible research designs: randomized controlled trials (RCTs), quasi-experimental (QED), single-subject or regression discontinuity. For example, one document's study was initially determined to be QED but provide scant description of variables by which the treatment and comparison group of teachers and their students were matched to be comparable. Another document failed because it reported a study that used an ex post facto (causal-comparative) research design and compared overall 4th and 6th grade scores at the district level from statewide assessments.

In Stage II of the screening process, two out of the 22 documents were eliminated from further consideration. One document had insufficient information to calculate an effect size. The other document was later found to focus much more on curriculum than on professional development.

Moreover, several documents were rejected during the post-coding stage when effect sizes were calculated using non-standard formulas. One document was found to be an earlier version of another, more complete report. After further review, a second document did not have sample size data to compute an effect size. A third document utilized hierarchical linear modeling as its quantitative analysis, but failed to provide sufficient information for the researchers to calculate a posttest effect size comparable with those from other documents. Finally a fourth document was eliminated after a series of homogeneity tests (see **Appendix C**) were run which showed that it had generated unusually large effect sizes.

Figure 3 also notes the inter-rater reliability at each stage of the coding process. The inter-rater reliability rate illustrates the degree of agreement between the assigned coders. As shown, the inter-rater reliability ranged from .81 to .95, showing a high degree of agreement between the two coders.

Results from the Coding Review

The coding and review process and the post-coding statistical analysis yielded 16 documents of studies to be included in the meta analysis, with two studies covering the same initiative, the Northeast Front Range Math/Science Partnership (MSP). The documents (from this point forward will be referred to as "studies") are listed in **Table 2**. Twelve studies reported on math professional development and student achievement effects and four studies reported on science. Six studies had randomized control trial or RCT designs, of which only one was in science. The other ten studies were conducted using a quasi-experimental design (QED) which requires matched treatment and comparison groups. Of those, three were on science, with the remainder dealing with math. Ten of the studies covered elementary grades (grades 1 through 6), seven covered middle grades (grades 7 and 8), and three reported results in the high school level.

Several types of student assessment instruments were used to generate measurable results for students across the studies. Eleven of the sixteen studies used at least one nationally known assessment or statewide standardized assessment. The remaining five relied on assessments specific to the professional development initiative and evaluation. The Lane study used released test items from the Colorado Student Assessment Program (CSAP), while the Jagielski study used released test items released from the National Assessment of Educational Progress (NAEP). Although on their own these are widely known standardized assessments, the use of specific test items from their respective pool suggests intent by the researcher to capture a specific phenomenon associated with the professional development initiative. Nine kinds of criterion-referenced instruments were used, including state assessments—Texas Assessment of Academic Skills (TAAS), Colorado, and Oregon Technology Enhanced State Assessment (TESA). Six national norm-referenced tests were employed in the studies—Metropolitan Achievement Test (MAT), the Iowa Test of Basic Skills (ITBS), ACCUPLACER, Terra Nova, and the Northwest Education Association Assessments (NWEA).

Six of the 16 studies relied on assessments that were unique to the project in order to measure student performance. These studies had small- to medium-sized groups of teachers participating in the professional development program, with a range of three teachers in one study to 87 in another. The number of assessed students varied from 63 to 936. Two studies aggregated student results to the classroom level, with one having 17 classes of students and another 20 classes. Ten of the studies utilized quasi-experimental designs (QED) that relied on comparable groups of teachers and students, while six studies had utilized random assignments of teachers to the treatment or control groups.

Table 2: List of Identified Studies and Key Study Characteristics

Study	Publication Status	Study Design	Content Area	School Level	Treatment Teachers N Size (All Teachers)	Treatment Students N Size (All Students)	Student Outcome Measure	Test Type/
Carpenter, et al., 1989	Journal Article	RCT	Math	Elementary	20 (40)	20 (40, by class)	ITBS (Level 7 Interviews on number facts & problem Study-specific tests (Scale 1,2,3)	National/Norm-referenced PD-specific/ PD-specific/
Dickson, 2002	Dissertation	QED	Science	Middle (8th) & High (9th & 10th)	4 (8)	86 (165)	Texas Assessment of Academic Skills (TAAS) (8th) End-of-Course Biology Test (9th & 10th)	State/Criterion-referenced State Norm-referenced
Heller et al., 2007	Report	RCT	Math	Elementary (2nd, 4th, 6th)	48	936 (1971)	Math Pathways and Pitfalls (MPP) Pitfalls Quiz	PD-specific
Jagielski, 1991	Dissertation	QED	Math	Elementary (3rd-6th), Middle (7th, 8th)	43 (70)	63 (70)	Study-specific assessment MCIP/89 using released NAEP test items	PD-specific/Criterion-referenced
Lane, 2003	Dissertation	QED	Math	Elementary	12 (22)	245 (490)	Constructed CSAP	PD-specific/Criterion-referenced
META Associates, 2006	Report	QED	Math	Middle (6th, 7th, 8th)	19 (34)	495 (767)	Colorado Student Assessment Program (CSAP)	State/Criterion-referenced
META Associates, 2007	Report	QED	Math	Middle 6th, 7th, 8th)	17 (40)	1099 (2256)	Student achievement as measured by Colorado Student Assessment Program (CSAP)	State/Criterion-referenced
Meyer & Sutton, 2006	Report	QED	Math	Middle (6th, 7th, 8th)	31 (155)	(7813)	Metropolitan Achievement Test (MAT) Criterion Referenced Test	Local/Criterion-referenced Local/Criterion-referenced
Niess, 2005	Report	RCT	Math	Elementary & Middle (3rd-8th)	24 (42)	310 (985)	Technology Enhanced State Assessment (TESA)	State/Criterion-referenced

Table 2 – continued

Study	Publication Status	Study Design	Content Area	School Level	Treatment Teachers N Size (All Teachers)	Treatment Students N Size (All Students)	Student Outcome Measure	Test Type/Reference
Palmer & Nelson, 2006*	Report	QED	Science	Elementary (5th, 6th), Middle (7th, 8th) & High (9th, 10th)	16 (43)	396 (792)	Northwest Evaluation Association (NWEA) assessments	National/Norm-referenced
Rubin & Norman, 1992	Journal Article	RCT	Science	Middle	7 (16)	108 (324)	Middle Grades Integrated Process Skill Test (MIPT)	PD-specific/Not reported
							Group Assessment of Logical Thinking Test (GALT)	Nat'l-Int'l/Criterion-referenced
Saxe, Gearhart, & Nasir, 2001	Journal Article	QED	Math	Elementary	17 (6)	17 (23, by class)	Study-specific assessments (Computational Scale) Study-specific assessments (Conceptual Scale)	PD-specific/Not reported PD-specific/Not reported
Scott, 2005	Dissertation	QED	Science	Elementary (3rd)	3 (6)	66 (100)	Iowa Test of Basic Skills (ITBS)	National/Norm-referenced
Siegle & McCoach, 2007	Journal Article	RCT	Math	Elementary (5th)	7 (15)	430 (872)	Math Achievement Test	National/Norm-referenced
Snippe, 1992	Report	RCT	Math	High	87 (198)	114 (274)	Terra Nova ACCUPLACER WorkKeys	National/Norm-referenced National/Norm-referenced National/Criterion-referenced
Walsh-Cavazos, 1994	Dissertation	QED	Math	Elementary (5th)	4 (6)	78 (111)	PSG Achievement Assessment	PD-specific/Not reported

Reporting and Analyzing Effect Size

An ***effect size*** (ES) is the "difference on a criterion measure between an experimental and a control group divided by the **control group's** standard deviation" (McMillan & Schumacher, 1997, p. 148). It provides a measure common across all the studies and gives a sense of the magnitude of the effect a treatment has on a dependent variable. For the CCSSO meta analysis study, we analyzed the effect teacher professional development has—in its various forms presented by the programs described in the studies—on student achievement outcomes (Lipsey & Wilson, 2001).

The sixteen studies generated a sum total of 104 effect sizes. **Table 3** reports several example effect sizes for each study as well as the range, and features the variety of effect sizes resulting from the many measures possible per study, including posttest only and pretest-posttest gains. The number of effects for each study ranged from two to 21 effects with an average of 6.5 effects per study. Six of the studies reported only two effect sizes. The Meyer and Sutton study reported ten effect sizes due to the abundance of test results generated from having three grades tested—grades 6 , 7 and 8—and from two types of tests that had several constructs such as concept and problem-solving, math procedures, algebra, computation, data analysis, geometry and measurement, and numeration. The Snippe study generated 21 effect sizes because all three standardized tests—Terra Nova, ACCUPLACER, and WorkKeys—were administered to six different study sites. The Jagielski study produced twenty effects as a result of comparisons of two treatment groups to the control group across five different test questions set according to NAEP proficiency levels. When analyzing multiple effects, we need to consider whether the effects are produced from independent samples of teachers and students. Dependence among effect sizes can arise when data are not drawn from independent samples.

To apply a meaning to the use of effect sizes for educators, one challenge is to translate the ES to something meaningful, e.g., practical effects, and *Cohen's d* statistic provides a useful guide (Lipsey & Wilson, 2001). Using the Cohen's *d* standard guidelines for effect sizes, 56 percent of the effect sizes in our study are small—0.01 to 0.2 is considered small. Twenty percent of the effect sizes were negative, suggesting that students of teachers who received the professional development treatment fared worse than their counterpart students. Nearly 8 percent of the 104 studies are considered to have small-medium or medium effect sizes, with medium set at $d = 0.5$. Only two effect sizes, one stemming from the Saxe et al. study (ES = 2.54) and another from the Snippe study (ES = .79) can be considered large ES, with five other studies coming close with ES ranging from .68 to .78. **Appendix B** provides a complete and detailed listing of all the effect sizes generated from each study.

Table 3: Example Effect Sizes Reported in Studies

Study	Number of Effects (Total=104)	Range of Effect Sizes	Student Outcome Measure	Effect Size	Cohen's d Standard
Carpenter, et al., 1989	7	0.11 to 0.69	Average posttest results from Iowa Test of Basic Skills (ITBS)	0.39	Small
			Interviews on number facts & problem	0.68	Medium
			Average across Scales 1-3 of study-specific test	0.32	Small
Dickson, 2002	2	0.10 to 0.43	Texas Assessment of Academic Skills (TAAS) (8^{th})	0.10	--
			End-of-Course Biology Test (9^{th} & 10^{th})	0.43	Small-medium
Heller et al., 2007	6	0.27 to 0.76	Pretest-posttest gain (4^{th}) on Math Pathways and Pitfalls (MPP) Pitfalls Quiz	0.69	Medium
Jagielski, 1991	20	-0.42 to 0.78	Average of pretest-posttest gains of both treatment groups on study-specific assessment-Level 250 NAEP test item	0.77	Medium-large
Lane, 2003	2	0.08 to 0.13	Pretest-posttest gain on Constructed CSAP	0.13	Small
META Associates, 2006	6	-1.52 to 0.22	Average of pretest-posttest gains (6^{th}, 7^{th}, 8^{th}) on Colorado Student Assessment Program (CSAP)	0.13	Small
META Associates, 2007	2	-0.19 to 0.11	Pretest-posttest gain on Colorado Student Assessment Program (CSAP)	-0.19	--
Meyer & Sutton, 2006	10	-0.10 to 0.13	Average of overall posttests (6^{th}, 7^{th}) in Metropolitan Achievement Test (MAT)	-0.02	--
			Overall posttest in Criterion Referenced Test	0.10	Small
Niess, 2005	4	-0.14 to 0.37	Pretest-posttest gain (Middle) in Technology Enhanced State Assessment (TESA)	0.11	Small
Palmer & Nelson, 2006	5	-0.21 to 0.11	Pretest-posttest gain (Grades 3^{rd}, 5^{th}, 6^{th}) in Northwest Evaluation Association (NWEA) assessments	0.11	Small
Rubin & Norman, 1992	8	-0.36 to 0.64	Pretest-posttest gain (Treatment vs. Control II) in Middle Grades Integrated Process Skill Test (MIPT)	0.64	Medium
			Pretest-posttest gain in (Treatment vs. Control II) Group Assessment of Logical Thinking Test (GALT)	0.12	Small
Saxe, Gearhart, & Nasir, 2001	6	-1.55 to 2.54	Average posttest results from study-specific assessments (Conceptual Scale)	1.63	Large
Scott, 2005	2	0.20 to 0.54	Pretest-posttest gain on Iowa Test of Basic Skills (ITBS)	0.20	Small
Siegle & McCoach, 2007	2	0.20 to 0.22	Cluster result on Math Achievement Test	0.20	Small
Snippe, 1992	21	-0.43 to 0.79	Terra Nova	-0.01	--
			ACCUPLACER	0.20	Small
			WorkKeys	.06	
Walsh-Cavazos, 1994	2	0.26 to 0.56	Pretest-posttest gain PSG Achievement Assessment	0.26	Small

Reviewing across studies, most of the effect sizes from the 16 studies are found to be modest. This often stems from controlling for prior testing results from both the treatment and comparison groups and examining if and by how much did students taught by teachers in the treatment group gain relative to their respective comparison group. For example, in the Niess (2005) study, students of teachers who participated in the professional development activities associated with the High Desert Math Science Partnership (MSP) initiative did improve on the state's standardized assessment (posttest ES = .13). However, after controlling for their prior performance on the assessment and comparing it to their counterparts whose teachers did not participate in the High Desert MSP initiative, the results remain positive but smaller (pretest-posttest gain ES = .10). A similar difference in effects between pre-post effect size and post-test effect size was found in the study results from Palmer and Nelson (2006), Meta Associates (2006), Lane (2003), Scott (2005), and Walsh-Cavazos (1994).

Another factor that may result in the modest effect sizes is the use of standardized assessments to capture student measurable outcomes as a result of the professional development initiatives. All the aforementioned studies used either statewide criterion-referenced assessments or nationally norm-referenced assessments. These tests may not be fine-tuned to capture the areas that the professional development initiatives are intending to impact. For example, the Lane study examined a professional development initiative with an objective of improving the problem-solving and reasoning skills of fifth grade students by deepening their teachers understanding of math concepts and providing them teaching strategies in problem solving and in modeling the use of questioning and critical thinking and new vocabulary to their students. The Colorado Student Assessment Program's standardized tests may not have captured the full measure of student gains in as a result of the professional development the students' teachers received.

Looking at it another way, studies that utilized student measures that are closer to the heart of what the professional development is intended to impact, do report larger effect sizes. In the Rubin and Norman study (1992), the researchers were evaluating a professional development initiative which trained middle school teachers in science processing skills and ways to model the science processing skills to their students. The study utilized the Middle Grades Integrated Process Skills Test (MIPT), a lesser-known assessment that measures student proficiency in understanding the skills with which scientists use to explore and analyze a phenomenon. Not surprisingly, the study found that students whose teachers participated in the professional development had greater understanding of the process skills compared to their non-equivalent counterparts whose teachers did not receive the professional development, even after controlling for prior performance (MIPT ES = .63). Similar cases can be found with the interview results from the Carpenter et al study (1989) with an ES of .68, Saxe, Gearhart and Nasir study (2001) with ES of 1.63 resulting from average posttest results on the conceptual scale of their study-specific assessment. Jagielski utilized released NAEP items for formulating her study-specific assessment, and the test items were selected as items to measure problem solving abilities of students with teachers who received (or did not receive) training in the problem solving standard from the NCTM Curriculum and Evaluation Standards for School Mathematics. Thus, it was not surprising to find that the pretest-posttest gain ES for the two treatment groups was .77 on one test item.

Two studies had more than one treatment group or comparison group. The Rubin and Norman study involved two control groups. The treatment group of teachers received professional development in the use of a systematic modeling strategy to increase the scientific approach process skills. The treatment group is compared to a control group of teachers that received training on a substitute strategy, the learning cycle. The second comparison group received no special training. Table 3 shows that students of teachers who received training in the use of the systematic modeling strategy exhibited a significant positive difference in their achievement process than their peers whose teachers received no special training.

The Jagielski study utilized a train-the-trainer model and thus involved two treatment groups. The first treatment group teachers attended problem-solving workshops at Loyola University and were trained by university staff. The second treatment group was composed of in-school colleagues recruited by teachers who attended the workshops. Both groups were compared against teachers who received no training. Table 3 shows that on average students of either treatment group exhibited a significant positive difference in their ability to understand the basic mathematical operations (addition, subtraction, multiplication, and division) and apply it in simple one-step word problems and in analyzing graphs and charts, as compared to their control counterparts.

Professional Development Features

The designs for providing professional development with teachers in the target, or treatment, group vary widely across the 16 studies in the meta analysis. It is possible to observe several patterns in the descriptive data for the set of professional development "programs" which typically include a combination of activities for improving teacher knowledge and skills. ***Content focus*** is not reported as a separate category in the table, but the content focus for teachers is consistently found in the descriptions of "Teacher Learning Goal." Content focus was a primary selection criterion for the meta analysis, and all the programs reported here sought to increase content knowledge of the teachers.

Table 4 displays the features by study and they varied considerably across the studies. First, the projects vary widely in ***time (contact hours)*** of professional development and ***duration*** (or overall period when implemented). Given that all of the studies reported did show positive effects on student achievement, we can see that there is an inconsistent pattern in the relationship of time and duration to effects. For example, the professional development initiatives included in the 16 studies are widely differing in total amount of time. One professional development design provided only two hours of further education for teachers, six studies reported less than 20 hours were devoted to teacher development, and four of the designs included a combination of activities totaling over 100 hours of teacher development. Current research shows that consistent effects are found when teachers have received over 100 hours of professional development (Banilower et al., 2006).

Table 4: Professional Development Features of the Studies

Study Authors, Year	Professional Development	Teacher Learning Goal	Teachers Location	PD Provider Agency	[illegible]	Months Duration	PD Components	Teacher Active Learning
Carpenter, et al., 1989	Cognitively Guided Instruction (CGI)	First grade teachers participate in a 4-week summer workshop to learn about research findings on learning and development of addition and subtraction concepts in young children and apply that learning in the classroom	24 schools in Madison (WI) metropolitan area	Researchers/ Authors	80	4.5	Summer institute Coursework In-service activity Study group Self-directed	Classroom mentoring Professional Network
Dickson, 2002	Inquiry Institute Science	K-12 teachers participate in an inquiry-based staff development program from "Immersion into Science" model (Loucks-Horsley et al, 1998)	Suburban school district north central Texas	School District	24	8	In-service Activity internship	Professional network
Heller et al., 2007	Mathematics Pathways and Pitfalls (MPP)	2nd-, 4th-, and 6th- grade teachers received introductory training and practice on strategies to motivate students to be critical thinkers of their math learning through logic and discourse	Five diverse districts across the U.S.	Researchers/ Authors	10	8	Summer institute In-service activity Internship	Lead instruction Observe
Jagielski, 1991	Mathematics Curriculum Improvement Project	Train-the trainer model, teachers receive training in problem-solving as recommended by the National Council of Teachers of Mathematics (NCTM) standards	Chicago, IL	University	36	8	In-service activity Conference Study group	Lead instruction Lead discussion Professional network
Lane, 2003	Problem-solving and reasoning Math	Improve 5th grade teachers knowledge of math concepts, problem solving, questioning & critical thinking, and new vocabulary	Five schools from the same school district in Colorado	Researcher/ Author	17	8	In-service activity Study group	Develop assessment Observe
META Associates, 2006	Northeast Front Range Math/Science Partnership (MSP)	Middle school math and science teachers participate in 2-week summer institutes, follow-up Saturday institutes and lesson study to gain content and pedagogical knowledge in geometry, earth/space science, force & motion, and/or life science	Five school districts in Colorado front range	Four Colorado universities science/math faculties and one science museum	120	7.5	Summer institute In-service activity Coaching Mentoring	Lead instruction Observe Develop assessment Professional network
META Associates, 2007	Northeast Front Range Math/Science Partnership (MSP)	Same as META Associates, 2006	Same as META Associates, 2006	Same as META Associates, 2006	120	7.5	Same as META Associates, 2006	Same as META Associates, 2006
Meyer & Sutton, 2006	Math in the Middle Institute Partnership	Train and support Grades 5-8 math teachers in math content knowledge enrichment, improved instructional strategies, and leadership skills	Lincoln, NE	University of Nebraska-Lincoln; Education Service Units	540	16	Summer institute In-service activity Courses	--
Niess, 2005	High Desert MSP Math teaching	Increase grades 3-8 math teachers' ability to teach the subject by enriching their content and pedagogical math knowledge, and incorporating collaborative techniques.	Five school districts in central Oregon	Oregon State University	304	8	Summer institute In-service activity	Professional network Lead instruction Observe

Table 4 - continued

Study Authors, Year	Professional Development	Teacher Learning Goal	Teachers Location	PD Provider Agency	[illegible]	Months Duration	PD Components	Teacher Active Learning
Palmer & Nelson, 2006	REC Lesson Study Science	For Grades 5-12 science teachers to increase content knowledge-- 2-week summer institute, improve pedagogy with Lesson Study, and apply new knowledge by designing lessons to present to class.	Ten school districts in Minnesota	MN university, Schwan Food, APEN Assoc, Global Education Resources	60	8	Summer institute Study group	Lead instruction Develop Assessment Observe Professional network
Rubin & Norman, 1992	Systematic Modeling Strategy Science Teaching	Train middle school teachers in science process skills and modeling teaching strategy for teaching science process skills to their students	Detroit, MI	Wayne State University	30	3	Courses In-service activity Mentoring	--
Saxe, Gearhart, & Nasir, 2001	Integrating Mathematics Assessment (IMA) or Collegial Support (SUPP)	IMA: Teacher learning focused on math concepts, understanding children's math, achievement motivations, integrated curriculum focus on fractions, measurement, & scale. Collaboration with other teachers interested in reformed (vs. traditional) instruction. SUPP: Teachers receive support and collaborative opportunities with others for implementing units on fractions, measurement & scale	Los Angeles metropolitan area	Researchers/ Authors	41	8	Summer institute In-service activity Study group Mentoring Internship	Lead instruction Develop assessment Observe Professional Network
Scott, 2005	TEAMS Professional Development Model	Build a community of professional learners, focus on instructional alignment via lesson studies, and established mentoring peer coaching through multiple activities and supports.	Suburban-Urban district Texas metropolitan area	School District	168	8	In-service activity Summer institute Conference Study group Coaching Mentoring	Professional network Lead discussion Classroom mentoring Observe
Siegle & McCoach, 2007	Self-Efficacy Teaching Strategies & Implementation Math	Train 5th grade math teachers in self-efficacy teaching strategies in 3 areas: 1) goal setting, 2) teacher feedback, 3) modeling followed by an implementation of measurement unit curriculum designed by the researchers	Ten districts varying urban, suburban, rural in six states (MA, MD, MI, MT, NC, NE)	University of Connecticut	2	1 day	In-service activity Coaching	Lead instruction Professional network
Snippe, 1992	National Research Center for Career and Technical Education (NRCCTE) model	Teams of career and technology education (CTE) and math teachers learn how to improve math instruction embedded in CTE curricula by team building, using curriculum maps aligned by math concept and CTE curricula, designing lesson plans that incorporate the NRCCTE model's seven elements.	Teachers from several states; providers traveled to each location	University of Minnesota	14	3 days	In-service activity Study group	Professional network Classroom mentoring
Walsh-Cavazos, 1994	Probability, Statistics, and Graphing (PSG) Module	Teachers participate in 12 hour training in PSG module, involving manipulatives, problem-solving, and concept-development techniques	South Texas school district	Researcher/ Author	12	3 days	In-service activity	--
				Mean Range	91 hrs. 2 - 540 hrs.	6 mos. 1day - 16mos.	3.3 activities 1 - 6 activities	2.1 types 1 - 4 types

The professional development designs reported in the 16 studies were carried out from 1990 to the present. The federal legislation and regulations under NCLB encouraged states and districts to plan teacher development for a given teacher to include more hours over a longer duration, which reflects the research studies of the 1990s. The studies reporting the largest number of hours of development time per teacher were carried out since 2000.

The providers of professional development in these studies are primarily from universities, and the researcher/evaluator producing the study is often from the same institution. It is likely that having access to evaluation expertise in a university is a major advantage for providers of professional development, and student achievement effects is likely enhanced by the professional development providers being with a university.

One key finding from Table 4 is the evidence of multiple professional development activities, ***follow-up steps*** with teachers in their schools, and ***active learning methods*** that were used with teachers. The descriptive information on the professional development provided in these programs that did have effects on improving student achievement show confirmation of evidence from prior research on the importance of continuing learning reinforcement activities after the initial period of teacher training or intensive knowledge development such as through a summer institute. These effective programs included from two to six different types of activities, including coaching, mentoring, internship, professional networks, and study group, in addition to coursework or initial in-service education. The meta analysis of studies was somewhat limited in being able to identify all activities that were carried out. But even so, the review procedures for the 16 studies produced strong evidence of active methods of teacher learning during professional development such as leading instruction, discussion with colleagues, observing other teachers and developing assessments, and professional networks.

Another key finding revealed in Table 4 is the nature of ***teacher learning goals*** in the professional development designs. Each of the brief descriptions shows clearly that these programs focused on helping teachers improve their knowledge of how students learn in the specific subject area, how to teach the subject with effective strategies, and the important connections between the subject content and appropriate pedagogy so that students will best learn. It is apparent that these programs were well planned to maximize the use of time with teachers so that the content of the professional development could be directly translated by the teacher into improvements in curriculum and instruction.

One finding from prior research was that effectiveness is improved with ***collective participation*** of teachers; that is, teachers are learning with others from their school or department. To maximize collective involvement of teachers, some designs focus on the whole school for teacher development—all teachers are part of the training and assistance. The set of studies in this analysis show mixed evidence of teachers' collective participation in the professional development. Several of the studies are clearly from programs focused at school-level (e.g., Dickson, 2002; Lane, 2003; Scott, 2005) and did involve teachers who are teaching in the same context and thus are learning together. But other study descriptions indicate that teachers traveled off-site, enrolled, or volunteered for the intensive initial content and pedagogy training period, which would mean less chance of collective participation in development with their teaching colleagues.

Results from Analysis: Common Findings Across Studies

With the total number of effect sizes identified across the 16 studies in our meta analysis we can examine the extent to which there are significant group differences. The results of the analysis of means are displayed in **Table 5**, separately for Mathematics and Science. Our analysis first categorized all the studies under mathematics or science and the method of measuring effect (pre-post analysis vs. post-analysis only). In the mathematics education studies that employed pre-post measures for determining effect size, a total of 21 effect sizes were reported and the mean effect size was .21. Among the math studies that used a post-test only method of measuring effects, a total of 68 effect sizes were reported and the mean effect size was .13. The table below summarizes the differences in means and number of studies by major research and measurement categories. We are focusing the analysis on mathematics. The number of effect sizes for science teacher professional development studies was small (pre-post: 10 effect sizes, post-analysis: 7 effect sizes) and the means for the effect sizes in each category were small and not significantly different from zero. See **Appendix C** for details on computation of effect sizes.

Studies that used randomized control trials (RCT) had significantly larger effect sizes than studies that were based on quasi experimental designs (QED) though both sets of studies also showed significant heterogeneity. For the pre-post studies, the mean effect size was .27 for those studies using random trials as compared to a mean of.17 for studies based on quasi experimental designs, which is a significant difference although the mean effect sizes are not substantively large (see Q values for both sets of math effects in Table 5a).

We also analyzed the mean effect sizes according to differences in the measures of student achievement that were used in the studies. Based on 15 effect sizes, the studies that used a pre-post test design and employed achievement measures that were aligned to the professional development treatment objectives (e.g., treatment focus on teaching geometric concepts and students are assessed on knowledge of geometric concepts) had a mean effect size of .32. Six effect sizes were found for studies that used statewide assessment results in mathematics as the outcome measure, and the mean effect size was only .01. Both of these sets of effects showed significant heterogeneity as well.

For the studies that used a post-analysis only (comparing outcomes between treatment and control groups of teachers), four types of achievement tests were found. The mean effect size for the 25 effects based on a program-specific student assessment was .28, a moderate average effect that is educationally meaningful. The mean for 25 effects based on national norm-referenced assessments was .17, a statistically significant result but a smaller effect size. The mean effect size for 11 studies that used local achievement tests was .05, a statistically significant finding but an average indicating less educational importance. The studies that used statewide criterion-referenced assessments had a small mean negative effect size (-.07) indicating no average positive effect and there was wide variation in effect sizes across the seven studies.

Table 5a: Mean Effect Sizes for Teacher Professional Development Effects On Student Achievement, Mathematics Studies

Categories	Math Pre-Post Mean Effect Size (SE)	N Effects	95 % CI	*Q* statistic	Math Post-Only Mean Effect Size (SE)	N Effects	95 % CI	*Q* statistic
Math Studies	0.21 (0.08)	21	(0.06, 0.36)	Q_T = 153.72*	0.13 (0.03)	68	(0.07, 0.20)	Q_T = 328.78*
Research Design				$Q_B(1)$ = 46.12*				$Q_B(1)$ = 66.72*
RCT	0.27 (0.13)	5	(0.01, 0.53)	Q_W = 53.24*	0.26 (0.05)	35	(0.16, 0.35)	Q_W = 78.37*
QED	0.17 (0.08)	16	(0.01, 0.34)	Q_W = 54.35*	0.04 (0.04)	33	(-0.04, 0.11)	Q_W = 183.70*
Measure Type				$Q_B(1)$ = 84.46				$Q_B(3)$ = 90.43*
PD Specific	0.32 (0.08)	15	(0.16, 0.49)	Q_W = 46.81	0.28 (0.09)	25	(0.10, 0.46)	Q_W = 91.73*
State Criterion- Referenced	0.01 (0.08)	6	(-0.15, 0.16)	Q_W = 22.45	-0.07 (0.14)	7	(-0.35, 0.21)	Q_W = 111.25*
National Norm-Referenced	--	--		--	0.17 (0.04)	25	(0.10, 0.24)	Q_W = 16.33
Local Test	--	--		--	0.05 (0.02)	11	(0.02, 0.09)	Q_W = 19.05*

N Effects = number of effect sizes per category (across studies identified with at least one significant effect size); *$p < .05$; if Q_T is significant a random-effects model is applied. If Q_W is not significant a fixed-effects model is applied. If Q_W is significant a random-effect model is used for that category. Q_B refers to differences between groups.

Table 5b: Mean Effect Sizes for Teacher Professional Development Effects On Student Achievement, Science Studies

Categories	Science Pre-Post Mean Effect Size (SE)	N Effects	95 % CI	*Q* statistic	Science Post-Only Mean Effect Size (SE)	N Effects	95 % CI	*Q* statistic
Science Studies	0.05 (0.08)	10	(-0.11, 0.20)	Q_T = 31.57*	0.18 (0.24)	7	(-0.29, 0.64)	Q_T = 84.15*
Research Design				$Q_B(1)$ = 1.36				$Q_B(1)$ = 33.23*
RCT	0.13 (0.20)	4	(-0.26, 0.53)	Q_W = 24.50*	-0.15 (0.28)	4	(-0.71, 0.41)	Q_W = 47.99*
QED	-0.02 (0.05)	6	(-0.12, 0.09)	Q_W = 5.71	0.63 (0.16)	3	(0.32, 0.94)	Q_W = 2.94
Measure Type				$Q_B(2)$ = 14.93*				$Q_B(3)$ = 47.27*
PD Specific	0.39 (0.23)	2	(-0.07, 0.85)	Q_W = 5.33*	0.12 (0.42)	2	(-0.71, 0.95)	Q_W = 17.41*
State Criterion- Referenced	--	--		--	0.67 (0.16)	2	(0.35, 0.98)	Q_W = 2.72
National Norm-Referenced	-0.02 (0.05)	6	(-0.12, 0.09)	Q_W = 5.71	0.54 (0.21)	1	(0.12, 0.96)	--
International	-.013 (0.24)	2	(-0.59, 0.34)	Q_W = 5.61*	-0.42 (0.42)	2	(-1.24, 0.40)	Q_W = 16.75*

N Effects = number of effect sizes per category (across studies identified with at least one significant effect size);*$p < .05$; if Q_T is significant a random-effects model is applied. If Q_W is not significant a fixed-effects model is applied. If Q_W is significant a random-effect model is used for that category. Q_B refers to differences between groups.

Professional Development Characteristics

We also conducted further analysis to examine any differences in mean effect sizes based on the grade span covered by the studies and any differences according to professional development design characteristics (see **Table 6**). We found that studies that targeted the elementary grades had larger mean effect sizes than studies that targeted middle school or high school grades. Fifteen effects from studies with the pre-post analysis design that covered elementary grades had a statistically significant mean effect of .32. With a post-only analysis design, thirty effects report a statistically significant mean effect size of .27. Furthermore, studies of professional programs that provide mentoring for participating teachers have a negative mean effect size of -.19, based on ten effects. Studies of programs that offer internships for their teachers have a positive mean effect size of .20 for nine effects. Based on studies with pre-post analysis design however, programs that offer collaborative networking for participating teachers show marginal (ES = .01, n = 6 effects) or near zero impact.

Studies with pre-post analyses design of programs had 15 effect sizes in which coherence was significant. Studies reporting two types of coherence have a mean effect size of .32 as contrasted to -.19 (none), .12 (one type), and -.00 (three types). Studies using a post-only analysis design had smaller effect sizes than those with pre-post analysis design. Post-only studies with two types of coherence report a consistently positive though smaller mean effect size (.14). According to research stemming from the Eisenhower study (Garet et al., 1999, 2001) and CCSSO's cross-state study (Blank et al., 2007), a professional development activity or program is more likely to be effective if it is a) consistent with the teacher's school curriculum or learning goals for students and/or aligned with state or district standards for student learning or performance, b) congruent to the day-to-day operations of schools and teachers, and c) compatible with the instructional practices and knowledge needed for the teachers' specific assignments. If the professional development program meets all three criteria and is aligned with overall policies and practices in the teacher's school system, then the professional development program helps undergird a supportive environment that encourages improvement in teaching practices and aids in the long-term sustainability of the changed practices (Grant, Peterson, & Shojgreen-Downer, 1996).

Table 6: Mean Effect Sizes and Certain Profession Development Designs and Characteristics, Mathematics Studies

Categories	Math Pre-Post Mean Effect Size (SE)	N Effects	95 % CI	*Q* statistic	Math Post-Only Mean Effect Size (SE)	N Effects	95 % CI	*Q* statistic
Grade Span				Q_B (1) = 84.46*				Q_B(2) = 71.24*
Elementary	0.32 (0.08)	15	(0.16, 0.49)	Q_W = 46.81*	0.27 (0.07)	30	(0.14, 0.41)	Q_W = 113.11*
Middle	0.01 (0.08)	6	(-0.15, 0.16)	Q_W = 22.45*	0.03 (0.04)	17	(-0.04, 0.10)	Q_W = 130.75*
High	--	--		--	0.11 (0.05)	21	(0.01, 0.22)	Q_W = 13.68
PD Design Components								
Receive Mentoring		--						Q_B(1) = 5.24*
Has Mentoring	--	--		--	-0.19 (0.24)	10	(-0.67, 0.28)	Q_W = 152.16*
None					0.16 (0.03)	58	(0.11, 0.22)	Q_W = 171.39*
Internship	--	--		--				Q_B(1) = 76.50*
Has Internship	--	--		--	0.21 (0.19)	9	(-0.16, 0.58)	Q_W = 76.12*
None	--	--		--	0.10 (0.03)	59	(0.04, 0.15)	Q_W = 176.17*
Collaborative Network (CB)				Q_B (1) = 84.46*	--	--		--
Has CB	0.01 (0.08)	6	(-0.15, 0.16)	Q_W = 22.45*	--	--		--
None	0.32 (0.08)	15	(0.16, 0.49)	Q_W = 46.81*	--	--		--
Active Learning				--				--
Develop Assessment or Review Student Work (DA)	--			--				Q_B(1) = 16.10*
Has DA					0.16 (0.03)	58	(0.11, 0.21)	Q_W = 171.30*
None					-0.20 (0.27)	10	(-0.72, 0.33)	Q_W = 141.38*
Coherence				Q_B(1) = 102.97*				Q_B(3) = 32.90*
None	-0.19 (0.04)	10	(-0.28, -0.11)	--	0.18 (0.04)	10	(0.11, 0.25)	Q_W = 9.65
1 Type	0.12 (0.08)	3	(-0.03, 0.27)	Q_W = .14	-0.43 (0.53)	3	(-1.47, 0.61)	Q_W = 81.07*
2 Types	0.32 (0.08)	15	(0.16, 0.49)	Q_W = 46.81*	0.14 (0.03)	53	(0.07, 0.20)	Q_W = 201.72*
3 Types	-0.00 (0.12)	2	(-0.24, 0.24)	Q_W = 3.80	0.23 (0.12)	2	(0.00, 0.46)	Q_W = 3.44

N Effects = number of effect sizes per category (across studies identified with at least one significant effect size); *$p < .05$; if Q_B is significant a random-effects model is applied. If Q_W is not significant a fixed-effects model is applied. If Q_W is significant a random-effect model is used for that category. Q_B refers to differences between groups.

Correlations of Professional Development Design Elements

Using the Pearson's product moment correlation statistic (r), we examined the data for any relationships between various elements of professional development (See **Appendix D** for full correlation table). Using a significance value of .01 (two-tail test), positive correlations were found among measures of time—contact hours, frequency and duration. In particular, statistically significant positive relationships were found to exist between total contact hours and frequency (r = .74), contact hours and duration (r = .83) and frequency and duration (r = .62). Among the types of professional development activities, statistically significant positive relationships exists between summer institute and contact hours (r = .577), and duration (r = .655), and for college courses and contact hours (r = .744) and duration (r = .596).

These findings confirm that professional development programs that involve summer institutes or courses for teachers also provide extensive time (through greater frequency, longer duration and more contact hours). Also, we found a statistically significant positive correlation between frequency and having two types or ways that the professional development programs are promoting coherence in teacher learning (r = .794). For example, High Desert MSP and Northeast Front Range MSP are geared not only toward teachers who need to acquire the "highly qualified" status under NCLB but are also designed so that students of participant teachers can meet state expectations for academic performance, as measured by their state assessments. Both of these programs provide over 100 hours for their participating teachers to learn and apply their learning through intensive summer institutes and follow-up activities during the school year.

In examining relationships between specific types of professional development activities and their means of actively engaging participant teachers in learning, statistically significant positive correlations were found between

- conference and leading a discussion (r = 1.000)
- summer institutes and developing assessments and reviewing student work (r = .345)
- summer institutes and observing other teachers (r = .418)
- study group and receive classroom mentoring (r = .579)
- classroom mentoring and engaging in learning network (r = .796 and
- classroom mentoring and developing assessments or reviewing student work (r = .883).

As examples, programs such as Integrated Mathematics Assessment (Saxe, Gearhart & Nasir, 2001) and Researchers in Every Classroom (Palmer & Nelson, 2006) are reported to actively engage teachers by providing them opportunities to observe other teachers and develop assessments or review their own students' work in summer institutes. Programs that incorporate study groups such as the NRCCTE model (Snippe, 1992) and Mathematics Curriculum Improvement Project (Jagielski, 1991) provide their participant teachers the opportunity to be actively engage through classroom mentoring and being part of a professional learning network. The data also show that when professional programs offer classroom mentoring, they are more than likely to engage those teachers in developing assessments and reviewing student work during those mentoring moments.

Summary of Findings

The CCSSO meta analysis of studies of teacher professional development programs in mathematics and science found that 16 studies reported significant effects of teacher development on improving student achievement. The evidence for the findings in the 16 studies were based on scientific research designs. These studies reported effect sizes for student achievement gains for a treatment group as compared to a control group and the studies provided adequate data and documentation for the CCSSO research team to compute or re-analyze effect sizes. The large majority (12 of 16) studies were focused on analyzing mathematics teacher professional development and effects on student achievement in mathematics. The mean effect size for mathematics studies using a pre-post design is 0.21. These results show consistent positive effect on gains in student achievement in mathematics from teacher professional development in mathematics education. The mean effect size for math studies using a posttest-only design is 0.13, indicating that student achievement is higher for students of teachers receiving professional development in math education than for students of comparable teachers who did not participate in professional development. Our meta-analysis identified four studies of professional development in science that had significant effects on student achievement.

The results for the 16 studies with effect sizes demonstrates to the education research and policy communities how meta analysis can and should be used in education to provide comparisons and aggregations of research findings over time and across many different studies. The process of review and analysis employed by CCSSO involved several thousand citations, initial pre-screening of 400 plus documents, and intensive coding and review of 74 studies. The methods of identifying, coding, and quantifying data used in the study can be employed for a variety of objectives in education research.

CCSSO reviewed the professional development program designs and learning goals documented in the 16 studies. We found several common patterns. The program designs included strong emphasis on teachers learning specific subject content as well as pedagogical content for how to teach the content to students. The implementation of professional development included multiple activities to provide follow-up reinforcement of learning, assistance with implementation, and support for teachers from mentors and colleagues in their schools. In terms of duration of development activities, 14 of the programs continued for six months or more. The mean contact time with teachers in program activities was 91 hours.

The numbers of teachers that were involved in the programs that were analyzed and found to be effective varied from less than ten to more than 90. The research and evaluation for the 16 studies employed multiple measures of student achievement and outcomes. The studies analysis of effects on student achievement included scales to measure learning in specific content areas (e.g., algebra, measurement). The use of multiple measures allowed use of different types of test items. A majority of the studies analyzed professional development for elementary and middle grades teachers. The analysis of effects showed a pattern of stronger effects for elementary level professional development than for middle or high school teachers.

Effect sizes were larger when measures of achievement were used that were specifically selected or developed to be aligned with the content focus of the professional development. However, the review of research did identify several studies with significant effects using large-scale statewide assessment programs. This result demonstrates to evaluators and decision-makers that professional development can be measured with readily available data thru annual student assessments. However the outcomes are not likely to appear as positive or consistent as an outcome measure specific to the treatment goals. Some studies that computed separate effect sizes by student grade, such as the Meta Associates 2006 study, showed that effects of professional development differed markedly by grade (e.g., posttest only results show strong positive effects in grade 6, negative effects in grade 7 and no effect in grade 8). Wide variation by grade may indicate that teachers' fidelity of implementation of their professional learning is related to the curriculum, or this kind of result may indicate differences in the content covered in student assessment instruments by grade.

One question that has been addressed in prior research is the effect of professional development on teachers and their knowledge and practice. The CCSSO meta analysis review did not include systematic identification or review of intervening measures of the professional development treatment, such as measures of gains in teacher knowledge, improvement in practices, or fidelity of implementation of what was learned. Several of the studies identified did report analysis of differences on these kinds of measures between teachers in the treatment and control groups. Further analysis across studies would provide stronger evidence and useful information about the relationship between professional learning of teachers from a specific initiative and subsequent improved learning by students.

The CCSSO meta analysis results show important cross-study evidence that teacher professional development in mathematics does have significant positive effects on student achievement. The analysis results also confirm the positive relationship to student outcomes of key characteristics of design of professional development programs that have been documented in prior research. The meta-analysis process and procedures carried out by CCSSO show strong potential for broader use and application in judging the validity and consistency of results across a range of education initiatives and the evidence of outcomes from the initiatives.

Meta-Analysis Results: How Findings Can Be Used by State Leaders

Based on the results of the meta-analysis of findings from teacher professional development studies, CCSSO can state several recommendations for how the results and processes from the meta-analysis can be useful to researchers, evaluators, and state education leaders.

- The meta-analysis design and procedures employed by CCSSO proved to be effective in identifying a set of common findings regarding effects of teacher professional development on student achievement, and the procedures proved useful to determine which studies and their results met high standards for scientific validity and reliability.
- A scientific research design can be efficiently employed to evaluate teacher professional development, and a design to measure effects of teacher development on subsequent student achievement should be strongly considered for each funded program for teacher and teaching improvement.

- The use of research designs involving treatment and control groups should become a regular practice and built into the plan and organization for professional development and other initiatives.
- Measures of implementation of professional development are critical to evaluation design in order to document and measure activities to reinforce and extend learning for teachers in their school setting.
- Multiple measures of student achievement should be included in the research design if possible to provide for different types of assessments of learning and analysis of subject content learned.
- State and local education data systems can be accessed by providers of professional development and evaluators and regular statewide or district-wide assessment instruments can be effective measures of outcomes.
- State leaders should ensure that data systems are structured so that data on teacher development initiatives can be linked to student achievement measures, and these data can be effective for evaluation even where individual identifiers are removed.
- Procedures for meta analysis modeled in this study provide a consistent, quantified methodology for application and use in other studies, including initial identification, multiple coding and validation of reviews, comparison of research design with established criteria, and consistent procedures for effect size analysis and coding of treatment variables.

Appendix A: Meta Analysis Coding Form Excerpt: Scaffolded Guide for Determining Inclusion of a Document

The document review process for the meta analysis study is aided by coding forms that coders and reconcilers complete in order to record systematically how they determined a document is included into the pool of studies to be analyzed. The systematic review is conducted with at least two coders in mind, with a third person as a reconciler. The coding forms are Excel file documents composed of multiple spreadsheets that 1) assist in determining whether a document in question is a candidate to be included in the analysis and 2) aid in the extraction of key data needed for the analysis. Each coder completes one form per document independently from his/her partner coder. The reconciler completes another similar form that combines the information from both coders by bringing their entered information side-by-side. The excerpt shown below is the first spreadsheet that guides a coder through the process of determining the viability of a document for entry into the pool of studies to analyze. At certain decision junctures, the coder is forced to consider whether the document should continue to the next round of reviews or should be rejected. A document could be rejected any of the decision points of the review process.

CCSSO, 2007 -- NSF Grant. No. REC-0635409

Coder's First Name:					
Unique Document No.:					
		Coded Fields	Coding Decision	Page No. /Section - Paragraph No.	Additional Instructions for coding decision
Bibliographic Information:					
1.	a. Author				[Text] Use APA-style for references. Example: Marek, E. A., & Methven, S. B. (1991). Effects of the learning cycle upon student and classroom teacher performance. *Journal of Research in Science Teaching. 28* (1), 41-53.
	b. Date				
	c. Report Title				
	d. City, State: Publisher/Institution				
Stage I Coding - Part I : Relevance of Document					
2.	Was the document published between Jan. 1, 1985 and August 31, 2007?				1=Yes, 0=No [Binary].
3.	Is the document reporting on a study that took place in the U.S. or its territories?				1=Yes, 0=No [Binary].
4.	Does the document report on study findings involving K-12 students and their teachers?	Specify grade level(s):			1=Yes, 0=No [Binary].
5.	a. Is the document from a journal, a book/book chapter, a thesis/dissertation or an unpublished report? If Yes, enter document type.				1=Yes, 0=No [Binary]. And [Text]. (Conference proceedings are acceptable only if they satisfy all other pertinent conditions in the coding form).
	b. Does the document contain an **empirical, quantitative study?** It should **NOT** be one of the following, which makes it ineligible for review: - literature review - research synthesis or meta-analysis - case study - qualitative study - commentary - opinion paper, or - theoretical paper (e.g., presenting a hypothesis or a model)?				1=Yes, 0=No [Binary].
6.	Does the document feature **in-service teacher professional development** program or a set of professional development activities for teachers?				1=Yes, 0=No [Binary]. Answer No if a) study is focused on pre-service teacher preparation, or b) study is focused on comprehensive reform models, curriculum, instructional models, teaching materials, assessments, or policies with little attention to professional development as a primary focus.

Appendix A continued

Coder's First Name:					
Unique Document No.:					
		Coded Fields	Coding Decision	Page No. /Section - Paragraph No.	Additional Instructions for coding decision
7.	Is the professional development in **math and/or science**? If Yes, enter subject.				[Responses=Math, Science, Both]. If **Both**, notify Nina at CCSSO and create another coding form with a new unique document number for the additional subject. There will be two coding analyses to reflect the two subjects covered.
8.	a. **Student achievement outcomes** in **mathematics**, may include: number sense, geometric concepts, algebraic concepts, measurement, data analysis, and logical reasoning.				1=Yes, 0=No [Binary].
	b. **Student achievement outcomes** in **science**, may include: knowledge in earth science, life science, and physical science, science inquiry skills, scientific reasoning, science experiment design, data interpretation and analysis, hypothesis testing, and explanation formulation from evidence.				1=Yes,0=No [Binary].
	c. Does the study provide at least one **student achievement outcome** in math or science as an **effect** of in-service teacher professional development?				1=Yes, 0=No [Binary]. .
9.	Does the study examine the effects of in-service professional development on **teacher outcomes** such as knowledge and skills, beliefs, and attitudes, and/or instructional practice?				1=Yes, 0=No [Binary].
10. Eligible designs:	a. Describe briefly the type of study and the study methods used. If manuscript is an empirical study, note all major components of the study.				[Text].
Randomized Controlled Trial (RCT)	b. Was random assignment used to place participants into different study groups? OR If a randomization procedure was not used, participants were placed into intervention groups using a process that was haphazard and functionally random? (see Appendix_Glossary)				1=Yes, 0=No [Binary]. An answer of "1" to either of these questions leads to a categorization of the study as a randomized controlled trial. However, the fact that haphazard assignment was used will be noted in the write-up of the intervention report. **If response is "1", skip Q c through e and go to Stage I-Part I Decision. If "0", go to the next question below.**

A document has to meet one of only four types of research designs to be considered for inclusion: randomized control trials, quasi experimental designs, single subject design or regression discontinuity. This is to guarantee that the document captures an empirical study.

Appendix A continued

Coder's First Name:					
Unique Document No.:					
		Coded Fields	Coding Decision	Page No. /Section - Paragraph No.	Additional Instructions for coding decision
Quasi-Experimental Design (QED)	c. Is the design a quasi-experimental design with EITHER 1) statistical controls for study participants' characteristics (e.g., teacher content knowledge, or student pre-test achievement measures), OR 2) comparison group(s) matched on study participants' characteristics.				1=Yes, 0=No [Binary]. An answer of "1" to either of these questions leads to a categorization of the study as a quasi-experimental design. **If response is "1", skip questions d through e and go to Stage I-Part I Decision. If "0", go to the next question.**
Single Subject Design (SSD)	d. Is the document's study a single-subject design?				1=Yes, 0=No [Binary]. A single subject design involves an individual subject whose behavior is observed for changes associated with the intervention or removal of treatment. **If response is "1", skip question e and go to Stage I-Part I Decision. If "0", go to the next question.**
Regression Discontinuity Design (RDD)	e. Is the document's study a regression discontinuity design?				1=Yes, 0=No [Binary]. A regression discontinuity design uses pretest-posttest program-comparison group strategy, but has the unique characteristic of assigning participants to program or comparison groups based solely on a cutoff score on a pre-program measure.
	Stage I - Part I Decision				1=Pass, 0=Fail. If marked "1" to all Q 2 through 8c AND one in Q 10, the document passes Part I. Proceed to next question. If not, mark as "0" to fail and stop – the document is ineligible for further review.
Stage I - Part II : Outcome Measures & Methodology					
11.	a. Describe the student ACHIEVEMENT outcome measures and constructs reflected in the outcome measures and the approach to measurement in Table 1a.				**Complete Table 1a. Please note that we are only concerned about student achievement outcomes. Therefore, outcome measures on student belief or attitudes should be excluded from Table 1a.**
	b. Only after passing Stage I-Part I Decision, does one or more of the student achievement outcome measures in Table 1 (1) have face validity, OR (2) report reliability, OR (3) is a standardized test?				1=Yes, 0=No [Binary].

The scaffolded guide spreadsheet is followed by additional spreadsheets whereby coders/reconcilers record data on a) student outcome measures and constructs (validity & reliability of those measures–Table 1a); b) teacher outcome measures and constructs (Table 1b); c) number of teachers participating in the study, by treatment and control groups (Figure 1a); d) number of students participating in the study by treatment and control groups (Figure 1b); e) teacher characteristics, by treatment and control groups (Table 2); f) characteristics of students, by treatment and control groups (Table 3); g) characteristics of the professional development initiative; and h) estimates of treatment effects (effect sizes), by outcome measures (Tables 5a-d).

Appendix A continued

Coder's First Name:					
Unique Document No.:					
		Coded Fields	Coding Decision	Page No. /Section - Paragraph No.	Additional instructions for coding decision
	c. Only after passing Stage I-Part I Decision, are effect sizes reported in the study? OR Is there any information that will allow computation of effect sizes for one or more of the measures above? For example — - **Mean, standard deviation** (or **standard error**), and sample size in each group - **Difference in means, pooled standard deviation** (or **standard error**), and **sample size** in each group - **Means, sample size**, and **t-value** (for two independent groups) - **Difference in means, sample size** and **p-value** - **Proportion** and **sample size** in each group - **Regression coefficients** -- ordinary least squares (**OLS**) or hierarchical linear modeling (**HLM**), **standard errors, degree of freedom**, and **sample size** in each group?				1=Yes, 0=No [Binary].
12.	Describe the teacher outcome measures and constructs reflected in the outcome measures and the approach to measurement in Table 1b.				**Complete Table 1b. Please include ALL teacher outcome measures including those on teacher content/pedagogical content knowledge, beliefs, and attitudes.**
	According to the reconciliation, the design of this study is:				
	Make sure to complete information for the applicable design.	**If RCT, go to Q 13. If QED, go to Q15. If Single-subject design, go to Q 16. If Regression discontinuity design, go to Q 17.**			
13. (RCT design only)	a. Describe specific details of the procedure of randomization or a procedure that was haphazard and functionally random.				[Text].
	b. Did the authors provide details of the randomization procedure in the document?				1=Yes, 0=No [Binary]. For an RCT to Meet Evidence Standards, the study participants (e.g., teachers, classrooms or students) should have been placed to each study condition through random assignment or a process that was haphazard and functionally random. If the assignment process in an RCT is truly random or functionally random as described above, the RCT
	c. Have the study participants (e.g., teachers or students) been placed to each study condition through random assignment or a process that was haphazard and functionally random?				

Appendix A continued

Coder's First Name:					
Unique Document No.:					
		Coded Fields	Coding Decision	Page No. /Section - Paragraph No.	Additional instructions for coding decision
	d. Is the study free of any other problems with randomization (e.g., subjects being replaced or switched between groups after initial random assignment)				Meets Evidence Standards, unless one or more of the following conditions is violated: (1) Randomization, (2) attrition, (3) teacher-intervention confound, and (4) intervention disruption.
	e. Does the RCT study have any **randomization problem**?				1=Yes, 0=No [Binary]. **Based on the responses in questions a-d, respond either "1" or "0".**
14. (RCT design only)	a. Describe how the author(s) addressed baseline equivalence for both student and teacher data. If there are any concerns of incomparability, describe them as well.				**Complete Table 2-** teachers' characteristics prior to professional development. In addition, **complete Table 3** - students' characteristics prior to their teachers' professional development. **Use "NR" in the major cells to mean that the data was not reported. Use "NA" to mean that the question and subsequent response(s) are not applicable.**
	b. Is Table 2 -Teacher's Characteristics completed?				1=Yes, 0=No [Binary].
	c. Is Table 3 - Student Characteristics completed?				1=Yes, 0=No [Binary].
	d. Is there incomparability in teacher baseline characteristics that is NOT corrected for in the impact estimates reported? Is so, please describe.				1=Yes, 0=No [Binary]. And [Text].
	e. Is there incomparability in student baseline characteristics that is NOT corrected for in the impact estimates reported? Is so, please describe.				1=Yes, 0=No [Binary]. And [Text]. **Skip to Q 18.**
15. (QED designs only)	a. Describe how the author(s) addressed baseline equivalence for both student and teacher data. If there are any concerns of incomparability, describe them as well.				**Complete Table 2** teachers' characteristics prior to professional development. In addition, **complete Table 3** - students' characteristics prior to their teachers' professional development.
	b. Is Table 2 -Teacher's Characteristics completed?				1=Yes, 0=No [Binary].
	c. Is Table 3 - Student Characteristics completed?				1=Yes, 0=No [Binary].

Appendix A continued

Coder's First Name:					
Unique Document No.:					
		Coded Fields	Coding Decision	Page No. /Section - Paragraph No.	Additional Instructions for coding decision
	d. Was equating (teaches and/or students) accomplished through **matching** involves creating or identifying intervention and comparison groups that "look" similar on a pretest of the outcome measure? (Criteria for matching may include some demographic variables)				1=Yes, 0=No [Binary].
	e. Was equating (teachers and/or students) accomplished through **statistical adjustment** involves using statistical procedures (e.g., covariate adjustment in an ANCOVA) to equate groups on pretest and address baseline incomparability in the impact analysis?				1=Yes, 0=No [Binary].
	d. Do the **teacher** (treatment and comparison) groups appear to be patently **incomparable** at baseline, and was the incomparability unlikely to be adequately addressed through statistical adjustment?				1=Yes,0=No [Binary]. If response is "1", this is an indication of **baseline equivalence problem.**
	e. Do the **student** (treatment and comparison) groups appear to be patently **incomparable** at baseline, and was the incomparability unlikely to be adequately addressed through statistical adjustment?				1=Yes, 0=No [Binary]. **Skip to Q 18.**
16. (Single subject designs only):	a. Was the sample size one?				1=Yes, 0=No [Binary].
	b. Was a single-subject design most appropriate or would a group design be a better option?				1=Yes, 0=No [Binary].
	c. Were the observation conditions standardized?				1=Yes, 0=No [Binary].
	d. Was the behavior that was observed defined operationally?				1=Yes, 0=No [Binary].
	e. Was the measurement highly reliable?				1=Yes, 0=No [Binary].
	f. Were sufficient repeated measures taken?				1=Yes, 0=No [Binary].
	g. Were the conditions in which the study was conducted described fully?				1=Yes, 0=No [Binary].
	h. Was there stability in the baseline condition before the treatment was introduced?				1=Yes, 0=No [Binary].
	I. Was there a difference between the length of time or number of observations between the baseline and the treatment conditions?				1=Yes, 0=No [Binary].
	j. Was only one variable changed during the treatment condition?				1=Yes, 0=No [Binary].

Appendix A continued

Coder's First Name:					
Unique Document No.:					
		Coded Fields	Coding Decision	Page No. /Section - Paragraph No.	Additional Instructions for coding decision
	k. Were threats to internal and external validity addressed?				1=Yes, 0=No [Binary]. **Skip to Q 18.**
17. (For regression discontinu-ity designs only):	a. Was the cut-off criterion followed without exception? Describe cut-off criterion.				1=Yes, 0=No [Binary]. And [Text].
	b. Does the pre-post distribution follow a polynomial function?				1=Yes, 0=No [Binary]. If the true pre-post relationship is logarithmic, exponential, or some other function, the model given below is unspecified and estimates of the effect of the PD are likely to be biased.
	c. Does the comparison group have pretest variance?				1=Yes, 0=No [Binary]. There must be a sufficient number of pretest values in the comparison group to enable adequate estimation of the true relationship (i.e., pre-post regression line) for that group.
	d. Do the treatment and comparison group come from a single continuous pretest distribution with the division between groups determined by the cut-off score?				1=Yes, 0=No [Binary]. Both groups must come from a single continuous pretest distribution with the division between groups determined by the cutoff. In some cases one might be able to find intact groups (e.g., two groups of patients from two different geographic locations) which serendipitously divide on some measure as to imply some cutoff. Such naturally discontinuous groups must be used with caution because of the greater likelihood that if they differed naturally at the cutoff prior to the program such a difference could reflect a selection bias.
	e. Is the PD program uniformly implemented to all recipients under the same conditions (duration, frequency, kind and sequence)?				1=Yes, 0=No [Binary].
18. (All designs)	a. Describe any teacher-intervention confound problems.				[Text].
	b. Does the study assign more than one teacher per condition?				1=Yes, 0=No [Binary]. A teacher-intervention confound occurs when only one teacher is assigned to each condition.(NO means that there is just one teacher per condition.) **If response is "1", skip the next two questions and go to Q 19.**

Appendix A continued

Coder's First Name:					
Unique Document No.:					
		Coded Fields	Coding Decision	Page No. /Section - Paragraph No.	Additional instructions for coding decision
	c. If there is only one teacher per condition, is there any evidence that teacher effects are negligible?				1=Yes, 0=No [Binary]. If response is "1" then go to the next question.
	d. Does the study have any teacher-intervention confound problem?				1=Yes, 0=No [Binary]. Answer "1" if "0" are the responses to the two prior questions.
19. (All designs)	a. Describe any overall or differential attrition problems, either reported by the authors or detected by the coders.				[Text]. Complete Figure 1 by providing the total number of participants (i.e., teachers and students) as well as the number of participants who dropped out of the study and/or the analysis. Make sure to indicate the unit of assignment or analysis, and specify the unit itself (e.g., student, teacher, class, or school).
	b. Is there any severe overall attrition problem in the study, either reported by the authors or detected by the coders?				1=Yes, 0=No [Binary]. Overall attrition is defined as a failure to measure the outcome variable on all the participants initially assigned to the intervention and comparison groups. (If a study begins with 100 students total and ends up with 79 students total: 79/100 = 0.79, then subtract from 1.0. Attrition is 1.0- 0.79 = 0.21, or 21%). Coders will determine on a case-by-case basis if there is a severe overall attrition problem in the study.
	c. Is there any severe differential attrition problem in the study, either reported by the authors or detected by the coders?				1=Yes, 0=No [Binary]. Differential attrition refers to the situation in which the percentage of the original study sample retained in the follow-up data collection is substantially different for the intervention and the comparison groups. Severe differentiation makes the results of a study suspect because it may compromise the comparability of the study groups. Coders will determine on a case-by-case basis if there is a severe differential attrition problem in the study.
	d. Did the author(s) present evidence of post-attrition group equivalence on pretest data (see instructions)?				1=Yes, 0=No [Binary]. If the authors did not report overall and differential attrition, they must present evidence of post-attrition group equivalence on pretest data. Post-attrition group equivalence on pretest data may be demonstrated by a well-powered (0.80) test of equivalence that is non-significant, or a standardized mean difference between groups of less than d=0.10.

Appendix A continued

Coder's First Name:					
Unique Document No.:					
		Coded Fields	Coding Decision	Page No. /Section - Paragraph No.	Additional Instructions for coding decision
	e. Does the study have any attrition problems?				1=Yes, 0=No [Binary]. Unless response to Q 19d is "1", code as "1" if Q 19b or Q 19c is "1". Otherwise, code as "0".
20. (All designs)	a. Describe any problem of disruption or contamination in intervention.				[Text]. Intervention contamination occurs when something happens after the beginning of the intervention and affects the outcome for the intervention or the comparison group, but not both. Describe any disruptions of the intervention or control condition, any contamination of the treatment group, or any contamination of the comparison group.
	b. Is there evidence of obvious disruption or intervention contamination that could have caused observed differences between the intervention and control groups?				1=Yes, 0=No [Binary]. Indication of problem with disruption or contamination in intervention
21. (All designs)	a. Were the unit of assignment and analysis described?				1=Yes, 0=No [Binary].
	b. Describe the unit of group assignment.				[Text].
	c. Describe the unit of analysis.				[Text].
	d. Does the unit of analysis match with the unit of assignment?				If there is a misalignment between unit of assignment and analysis, clustering corrections should be made. Notify Nina at CCSSO about the this.
22. (All designs)	a. Was there any serious violations of statistical assumptions or any serious bias in reporting of findings?				1=Yes, 0=No. [Binary]. If responding as "1" go to next question. If "0", skip the next question and go to Q 23.
	b. Describe any serious violation of statistical assumptions or bias in reporting or findings.				[Text].
	SUM of the number of problems in randomization, baseline equivalence, attrition, teacher-intervention confound, or disruption of intervention (from min of 0 to max of 4)		0		The default starting value is zero (0).

Appendix A continued

Coder's First Name:					
Unique Document No.:					
		Coded Fields	Coding Decision	Page No. /Section - Paragraph No.	Additional Instructions for coding decision
	Stage I - Part II Decision				1=Pass, 0=Fail [Binary]. The document passes and goes on for further review, IF: - As an **RCT**, it has **NO or ONE problem in** randomization, attrition, teacher-intervention confound, or disruption. - As a **QED**, it has **NO** problem in baseline equivalence, attrition, teacher-intervention confound, or disruption. - As an **SS** design, it met **ALL** the conditions - As an **RD** design, it met **ALL** the conditions If the document fails, stop -- the document is ineligible for further review.
Stage II Coding : Documentation of Effect Sizes and PD Features					
23. (Effect sizes)	a. Document overall and subgroup means, standard deviation (SD), and N size for both treatment and comparison groups, and the time of measurement (e.g. pretest, posttest, follow-up test). Are **Tables 5a, b, c, and d** completed?				1=Yes, 0=No. [Binary]. Complete table **5a** to enter outcome measures that are based on **continuous variables**; complete table **5b** to enter outcomes that are based on **dichotomous variables** (e.g., percent proficiency); complete table **5c** to make Benjamin Hochberg corrections and table **5d** for clustering corrections.
	b. Do the effect sizes need to be computed using non-standard formulas?				1=Yes,0=No. [Binary]. If Yes, set aside. Notify Nina at CCSSO for assistance in computing effect sizes.
	c. Do results pertain to multiple periods of follow ups beyond the post-test?				1=Yes,0=No. [Binary].
24. (PD features)	a. Document the characteristics of the professional development intervention. Is **Table 4** completed?				1=Yes,0=No. [Binary]. **Complete table 4.**
	b. Based on the information provided on the content and implementation of the professional development, is there enough information to facilitate replication of the intervention?				1=Yes,0=No. [Binary].

Appendix A continued

Coder's First Name:					
Unique Document No.:					
		Coded Fields	Coding Decision	Page No. /Section - Paragraph No.	Additional Instructions for coding decision
	c. Is/are the author(s) of the document the evaluator of the intervention, OR designer of the intervention, AND/OR implementer of the intervention? If Yes, enter type(s).				1=Yes,0=No. [Binary]. And [Text].
25.	Was the effect of the professional development on student achievement confounded with the effect of curriculum?				1=Yes, 0=No. [Binary]. Often it is difficult to disentangle the effect of professional development on student achievement from the effect of related curriculum if they are interwoven in the PD activity.
26.	Do the measures for student outcomes align with the professional development?				1=Yes, 0=No. [Binary]. Misalignment between the student outcome measures and professional development introduces analytic complexities and limits interpretations of results.
	Additional Comments				[Text]. Add any other information that will assist in capturing the nature of the study design, measures, outcome results, and/or context.
	Stage II Status				**1=Completed, 0=To Be Determined/In Progress [Binary].**

Additional information about the coding form can be found in
http://www.ccsso.org/projects/improving_evaluation_of_professional_development/Meta_Analysis_Study/

Appendix B: Effects of Professional Development on Student Achievement, by Study (N = 104)

Study (No. of Effects)	Study Design	Outcome Measure	Unit of Analysis	Time of Measurement /Group	Effect Size	Cohen's d Standard	Posttest Only or Pretest-posttest Gain	Applied correction for clustering or multiple comparisons?
Carpenter, et al., 1989 **(7)**	RCT	ITBS (Level 7), Computation	By teacher group (treatment vs. comparison)	Posttest (1st & 2nd)	.41	Small-medium	Posttest only	Used adjusted mean
		ITBS (Level 7), Problems	By teacher group (treatment vs. comparison)	Posttest (1st & 2nd)	.37	Small	Posttest only	Used adjusted mean
		Interviews on number facts	By teacher group (treatment vs. comparison)	Posttest (1st & 2nd)	.66	Medium	Posttest only	Used adjusted mean
		Interviews on problem solving	By teacher group (treatment vs. comparison)	Posttest (1st & 2nd)	.69	Medium	Posttest only	Used adjusted mean
		Study-specific test, Simple Addition & Subtraction	By teacher group (treatment vs. comparison)	Posttest (1st & 2nd)	.43	Small-medium	Posttest only	Used adjusted mean
		Study-specific test ,Complex Addition & Subtraction	By teacher group (treatment vs. comparison)	Posttest (1st & 2nd)	.42	Small-medium	Posttest only	Used adjusted mean
		Study-specific test, Advanced Word Problems	By teacher group (treatment vs. comparison)	Posttest (1st & 2nd)	.11	Small	Posttest only	Used adjusted mean
Dickson, 2002 (2)	QED	Texas Assessment of Academic Skills (TAAS) (8th)	By teacher group (treatment vs. comparison)	Posttest (Middle, 8th)	.096608	Small	Posttest only	No
		End-of-Course Biology Test (9th & 10th)	By teacher group (treatment vs. comparison)	Posttest (High, 9th-10th)	.43029	Small-medium	Posttest only	No
Heller et al., 2007 (6)	RCT	Math Pathways and Pitfalls (MPP) Pitfalls Quiz, Overall	By teacher/class	Posttest (2nd)	.41065	Small-medium	Posttest only	Yes
		Math Pathways and Pitfalls (MPP) Pitfalls Quiz, Overall	By teacher/class	Elementary (2nd)	.41241	Small-medium	Pretest-Posttest Gain	Yes
		Math Pathways and Pitfalls (MPP) Pitfalls Quiz, Overall	By teacher/class	Posttest (4th)	0.763156	Medium-large	Posttest only	Yes

Appendix B continued

Study (No. of Effects)	Study Design	Outcome Measure	Unit of Analysis	Time of Measurement /Group	Effect Size	Cohen's d Standard	Posttest Only or Pretest-posttest Gain	Applied correction for clustering or multiple comparisons?
		Math Pathways and Pitfalls (MPP) Pitfalls Quiz, Overall	By teacher/class	Elementary (4th)	.685868	Medium-large	Pretest-Posttest Gain	Yes
		Math Pathways and Pitfalls (MPP) Pitfalls Quiz, Overall	By teacher/class	Posttest (6th)	.352674	Small	Posttest only	Yes
		Math Pathways and Pitfalls (MPP) Pitfalls Quiz, Overall	By teacher/class	Elementary (6th)	.271791	Small	Pretest-Posttest Gain	Yes
Jagielski, 1991 (20)	QED	Study-specific assessment MCIP/89, NAEP Level 250-Question 1	By class	Posttest (3rd-8th) Treatment I vs. Control)	.256549	Small	Posttest only	Yes
		Study-specific assessment MCIP/89, NAEP Level 250-Question 1	By class	Treatment I vs. Control	.746684	Medium-large	Pretest-Posttest Gain	Yes
		Study-specific assessment MCIP/89, NAEP Level 250-Question 1	By class	Posttest (3rd-8th) Treatment II vs. Control)	.207456	Small	Posttest only	Yes
		Study-specific assessment MCIP/89, NAEP Level 250-Question 1	By class	Treatment II vs. Control	.784691	Medium-large	Pretest-Posttest Gain	Yes
		Study-specific assessment MCIP/89, NAEP Level 300-Question 2	By class	Posttest (3rd-8th) Treatment I vs. Control)	.40038	Small-medium	Posttest only	Yes
		Study-specific assessment MCIP/89, NAEP Level 300-Question 2	By class	Treatment I vs. Control	.546542	Medium	Pretest-Posttest Gain	Yes
		Study-specific assessment MCIP/89, NAEP Level 300-Question 2	By class	Posttest (3rd-8th) Treatment II vs. Control)	.057441	Medium	Posttest only	Yes
		Study-specific assessment MCIP/89, NAEP Level 300-Question 2	By class	Treatment II vs. Control	.366257	Small	Pretest-Posttest Gain	Yes
		Study-specific assessment MCIP/89, NAEP Level 350-Question 3	By class	Posttest (3rd-8th) Treatment I vs. Control)	.274124	Small	Posttest only	Yes
		Study-specific assessment MCIP/89, NAEP Level 350-Question 3	By class	Treatment I vs. Control	.20929	Small	Pretest-Posttest Gain	Yes

Appendix B continued

Study (No. of Effects)	Study Design	Outcome Measure	Unit of Analysis	Time of Measurement /Group	Effect Size	Cohen's d Standard	Posttest Only or Pretest-posttest Gain	Applied correction for clustering or multiple comparisons?
		Study-specific assessment MCIP/89, NAEP Level 350-Question 3	By class	Posttest (3rd-8th) Treatment II vs. Control)	.159811	Small	Posttest only	Yes
		Study-specific assessment MCIP/89, NAEP Level 350-Question 3	By class	Treatment II vs. Control	.137631	Small	Pretest-Posttest Gain	Yes
		Study-specific assessment MCIP/89, NAEP Level 350-Question 4	By class	Posttest (3rd-8th) Treatment I vs. Control)	.396558	Small	Posttest only	Yes
		Study-specific assessment MCIP/89, NAEP Level 350-Question 4	By class	Treatment I vs. Control	.252577	Small	Pretest-Posttest Gain	Yes
		Study-specific assessment MCIP/89, NAEP Level 350-Question 4	By class	Posttest (3rd-8th) Treatment II vs. Control)	.259288	Small	Posttest only	Yes
		Study-specific assessment MCIP/89, NAEP Level 350-Question 4	By class	Treatment II vs. Control	.664996	Medium	Pretest-Posttest Gain	Yes
		Study-specific assessment MCIP/89, Question 5	By class	Posttest (3rd-8th) Treatment I vs. Control)	.058814	Small	Posttest only	Yes
		Study-specific assessment MCIP/89, Question 5	By class	Treatment I vs. Control	-.42439	--	Pretest-Posttest Gain	Yes
		Study-specific assessment MCIP/89, Question 5	By class	Posttest (3rd-8th) Treatment II vs. Control)	-.26524	--	Posttest only	Yes
		Study-specific assessment MCIP/89, Question 5	By class	Treatment II vs. Control	-.41516	--	Pretest-Posttest Gain	Yes
Lane, 2003 (2)	QED	Constructed CSAP, Overall	By teacher group (treatment vs. comparison)	Posttest (Elementary)	.08	Small	Posttest only	No
		Constructed CSAP, Overall	By teacher group (treatment vs. comparison)	Elementary	0.126908	Small	Pretest-Posttest Gain	Yes

Appendix B continued

Study (No. of Effects)	Study Design	Outcome Measure	Unit of Analysis	Time of Measurement /Group	Effect Size	Cohen's d Standard	Posttest Only or Pretest-posttest Gain	Applied correction for clustering or multiple comparisons?
META Associates, 2006 (5)	QED	Colorado Student Assessment Program (CSAP)	By teacher group (treatment vs. comparison)	Posttest (Middle, 6th)	.22	Small	Posttest only	No
		Colorado Student Assessment Program (CSAP)	By teacher group (treatment vs. comparison)	Posttest (Middle, 7th)	-1.52	--	Posttest only	No
		Colorado Student Assessment Program (CSAP)	By teacher group (treatment vs. comparison)	Posttest (Middle, 8th)	0	None	Posttest only	No
		Colorado Student Assessment Program (CSAP)	By teacher group (treatment vs. comparison)	Grade 6th	.0864699	Small	Pretest-Posttest Gain	No
		Colorado Student Assessment Program (CSAP)	By teacher group (treatment vs. comparison)	Grade 7th	.1470775	Small	Pretest-Posttest Gain	No
		Colorado Student Assessment Program (CSAP)	By teacher group (treatment vs. comparison)	Grade 8th	.1435162	Small	Pretest-Posttest Gain	No
META Associates, 2007 (2)	QED	Student achievement as measured by Colorado Student Assessment Program (CSAP), Overall	By teacher group (treatment vs. comparison)	Posttest 2006 (Elementary & Middle, 4th-8th)	.110911	Small	Posttest only	Yes
		Student achievement as measured by Colorado Student Assessment Program (CSAP), Overall	By teacher group (treatment vs. comparison)	Elementary & Middle, 4th-8th	-.1933	--	Pretest-Posttest Gain	Yes
Meyer & Sutton, 2006 (8)	QED	Metropolitan Achievement Test (MAT) , Overall	By teacher group (treatment vs. comparison)	Posttest (Elementary,5th)	.023587	Small	Posttest only	No
		Metropolitan Achievement Test (MAT), Math Concepts & Problem Solving	By teacher group (treatment vs. comparison)	Posttest (Middle, 6th)	.074428	Small	Posttest only	No

Appendix B continued

Study (No. of Effects)	Study Design	Outcome Measure	Unit of Analysis	Time of Measurement /Group	Effect Size	Cohen's d Standard	Posttest Only or Pretest-posttest Gain	Applied correction for clustering or multiple comparisons?
		Metropolitan Achievement Test (MAT), Math Procedures	By teacher group (treatment vs. comparison	Posttest (Middle, 6th)	.045459	Small	Posttest only	No
		Metropolitan Achievement Test (MAT), Overall	By teacher group (treatment vs. comparison	Posttest (Middle, 6th)	.068535	Small	Posttest only	No
		Metropolitan Achievement Test (MAT), Overall	By teacher group (treatment vs. comparison)	Posttest (Middle, 7th)	-.09989	--	Posttest only	No
		Criterion Referenced Test, Overall	By teacher group (treatment vs. comparison)	Posttest (Middle, 8th)	.100606	Small	Posttest only	No
		Criterion Referenced Test, Algebra	By teacher group (treatment vs. comparison)	Posttest (Middle, 8th)	.124888	Small	Posttest only	No
		Criterion Referenced Test, Computation	By teacher group (treatment vs. comparison)	Posttest (Middle, 8th)	.027889	Small	Posttest only	No
		Criterion Referenced Test, Data Analysis	By teacher group (treatment vs. comparison)	Posttest (Middle, 8th)	.040299	Small	Posttest only	No
		Criterion Referenced Test, Geometry & Measurement	By teacher group (treatment vs. comparison)	Posttest (Middle, 8th)	.126806	Small	Posttest only	No
		Criterion Referenced Test, Numeration	By teacher group (treatment vs. comparison)	Posttest (Middle, 8th)	.048704	Small	Posttest only	No
Niess, 2005 (4)	RCT	Technology Enhanced State Assessment (TESA), Math Computation, Problem-Solving Skills	By teacher group (treatment vs. comparison)	Posttest (Elementary)	.362457	Small	Posttest only	No
		Technology Enhanced State Assessment (TESA), Math Computation, Problem-Solving Skills	By teacher group (treatment vs. comparison)	Elementary	-.1393	--	Pretest-Posttest Gain	No

Appendix B continued

Study (No. of Effects)	Study Design	Outcome Measure	Unit of Analysis	Time of Measurement /Group	Effect Size	Cohen's d Standard	Posttest Only or Pretest-posttest Gain	Applied correction for clustering or multiple comparisons?
		Metropolitan Achievement Test (MAT), Math Procedures	By teacher group (treatment vs. comparison	Posttest (Middle, 6th)	.045459	Small	Posttest only	No
		Metropolitan Achievement Test (MAT), Overall	By teacher group (treatment vs. comparison	Posttest (Middle, 6th)	.068535	Small	Posttest only	No
		Metropolitan Achievement Test (MAT), Overall	By teacher group (treatment vs. comparison)	Posttest (Middle, 7th)	-.09989	--	Posttest only	No
		Criterion Referenced Test, Overall	By teacher group (treatment vs. comparison)	Posttest (Middle, 8th)	.100606	Small	Posttest only	No
		Criterion Referenced Test, Algebra	By teacher group (treatment vs. comparison)	Posttest (Middle, 8th)	.124888	Small	Posttest only	No
		Criterion Referenced Test, Computation	By teacher group (treatment vs. comparison)	Posttest (Middle, 8th)	.027889	Small	Posttest only	No
		Criterion Referenced Test, Data Analysis	By teacher group (treatment vs. comparison)	Posttest (Middle, 8th)	.040299	Small	Posttest only	No
		Criterion Referenced Test, Geometry & Measurement	By teacher group (treatment vs. comparison)	Posttest (Middle, 8th)	.126806	Small	Posttest only	No
		Criterion Referenced Test, Numeration	By teacher group (treatment vs. comparison)	Posttest (Middle, 8th)	.048704	Small	Posttest only	No
Niess, 2005 (4)	RCT	Technology Enhanced State Assessment (TESA), Math Computation, Problem-Solving Skills	By teacher group (treatment vs. comparison)	Posttest (Elementary)	.362457	Small	Posttest only	No
		Technology Enhanced State Assessment (TESA), Math Computation, Problem-Solving Skills	By teacher group (treatment vs. comparison)	Elementary	-.1393	--	Pretest-Posttest Gain	No

Appendix B continued

Study (No. of Effects)	Study Design	Outcome Measure	Unit of Analysis	Time of Measurement /Group	Effect Size	Cohen's d Standard	Posttest Only or Pretest-posttest Gain	Applied correction for clustering or multiple comparisons?
		Technology Enhanced State Assessment (TESA), Math Computation, Problem-Solving Skills	By teacher group (treatment vs. comparison)	Posttest (Middle)	.128815	Small	Posttest only	No
		Technology Enhanced State Assessment (TESA), Math Computation, Problem-Solving Skills	By teacher group (treatment vs. comparison)	Middle	.105168	Small	Pretest-Posttest Gain	No
Palmer & Nelson, 2006 (5)	QED	Northwest Evaluation Association (NWEA) assessments, General Science	By teacher group (treatment vs. comparison)	Elementary (3rd, 5th, 6th)	.11	Small	Pretest-Posttest Gain	No
		Northwest Evaluation Association (NWEA) assessments, General Science	By teacher group (treatment vs. comparison)	Middle (7th, 8th)	.06	Small	Pretest-Posttest Gain	No
		Northwest Evaluation Association (NWEA) assessments, General Science	By teacher group (treatment vs. comparison)	High (9th, 10th)	-.21	--	Pretest-Posttest Gain	No
		Northwest Evaluation Association (NWEA) assessments, Inquiry	By teacher group (treatment vs. comparison)	Elementary (3rd, 5th, 6th)	-.01	--	Pretest-Posttest Gain	No
		Northwest Evaluation Association (NWEA) assessments, General Science, Inquiry	By teacher group (treatment vs. comparison)	High (9th, 10th)	-.11	--	Pretest-Posttest Gain	No
Rubin & Norman, 1992 (8)	RCT	Middle Grades Integrated Process Skill Test (MIPT)	By teacher group (treatment vs. comparison)	Posttest (Middle 6th-9th, Treatment vs. Control I)	-.29421	--	Posttest only	Yes
		Middle Grades Integrated Process Skill Test (MIPT)	By teacher group (treatment vs. comparison)	Posttest (Middle 6th-9th,Treatment vs. Control II)	.553343	Medium	Posttest only	Yes
		Middle Grades Integrated Process Skill Test (MIPT)	By teacher group (treatment vs. comparison)	Middle 6th-9th, Treatment vs. Control I	.165492	Small	Pretest-Posttest Gain	Yes

Appendix B continued

Study (No. of Effects)	Study Design	Outcome Measure	Unit of Analysis	Time of Measurement /Group	Effect Size	Cohen's d Standard	Posttest Only or Pretest-posttest Gain	Applied correction for clustering or multiple comparisons?
		Middle Grades Integrated Process Skill Test (MIPT)	By teacher group (treatment vs. comparison)	Middle 6th-9th, Treatment vs. Control II	.635319	Medium	Pretest-Posttest Gain	Yes
		Group Assessment of Logical Thinking Test (GALT)	By teacher group (treatment vs. comparison)	Posttest (Middle 6th-9th, Treatment vs. Control I)	-0.83405	--	Posttest only	Yes
		Group Assessment of Logical Thinking Test (GALT)	By teacher group (treatment vs. comparison)	Posttest (Middle 6th-9th, Treatment vs. Control II)	0	None	Posttest only	Yes
		Group Assessment of Logical Thinking Test (GALT)	By teacher group (treatment vs. comparison)	Middle 6th-9th, Treatment vs. Control I	-.35745	Small	Pretest-Posttest Gain	Yes
		Group Assessment of Logical Thinking Test (GALT)	By teacher group (treatment vs. comparison)	Middle 6th-9th, Treatment vs. Control II	.119162	Small	Pretest-Posttest Gain	Yes
Saxe, Gearhart, & Nasir, 2001 (6)	QED	Study-specific assessments, Computational Scale)	By teacher/class	Posttest (Elementary-Treatment II vs. Control)	-1.36	--	Posttest only	No
		Study-specific assessments, Computational Scale)	By teacher/class	Posttest (Elementary-Treatment I vs. Control)	-.55	--	Posttest only	No
		Study-specific assessments, Conceptual Scale	By teacher/class	Posttest (Elementary-Treatment II vs. Control)	.72	Medium-Large	Posttest only	No
		Study-specific assessments, Conceptual Scale	By teacher/class	Posttest (Elementary-Treatment I vs. Control)	2.54	Large	Posttest only	No
		Study-specific assessments, Overall	By teacher/class	Posttest (Elementary-Treatment I vs. Control)	-.5667	--	Posttest only	Yes
		Study-specific assessments, Overall	By teacher/class	Posttest (Elementary-Treatment II vs. Control)	-1.5541	--	Posttest only	Yes
Scott, 2005 (2)	QED	Iowa Test of Basic Skills (ITBS), Overall	By teacher group (treatment vs. comparison)	Posttest (3rd)	.542299	Medium	Posttest only	No
		Iowa Test of Basic Skills (ITBS), Overall	By teacher group (treatment vs. comparison)	Elementary (3rd)	.198872	Small	Pretest-Posttest Gain	No

Appendix B continued

Study (No. of Effects)	Study Design	Outcome Measure	Unit of Analysis	Time of Measurement /Group	Effect Size	Cohen's d Standard	Posttest Only or Pretest-posttest Gain	Applied correction for clustering or multiple comparisons?
Siegle & McCoach, 2007 (2)	RCT	Math Achievement Test	By school (treatment vs. comparison)	Posttest (Elementary, 5th, cluster)	.1959	Small	Posttest only	Yes
		Math Achievement Test	By school (treatment vs. comparison)	Posttest (Elementary, 5th, single site)	.2159	Small	Posttest only	Yes
Snippe, 1992 (21)	RCT	Terra Nova, Overall	By class	Posttest (High)	-.01	--	Posttest only	No
		Terra Nova	By class	Posttest (High, Site A)	-.43	--	Posttest only	No
		Terra Nova	By class	Posttest (High, Site B)	.15	Small	Posttest only	No
		Terra Nova	By class	Posttest (High, Site C)	.01	Small	Posttest only	No
		Terra Nova	By class	Posttest (High, Site D)	.13	Small	Posttest only	No
		Terra Nova	By class	Posttest (High, Site E)	.14	Small	Posttest only	No
		Terra Nova	By class	Posttest (High, Site F)	.04	Small	Posttest only	No
		ACCUPLACER, Overall	By class	Posttest (High)	.20	Small	Posttest only	No
		ACCUPLACER	By class	Posttest (High, Site A)	.3	Small	Posttest only	No
		ACCUPLACER	By class	Posttest (High, Site B	.03	Small	Posttest only	No
		ACCUPLACER	By class	Posttest (High, Site C	.45	Small-medium	Posttest only	No
		ACCUPLACER	By class	Posttest (High, Site D	.14	Small	Posttest only	No
		ACCUPLACER	By class	Posttest (High, Site E	-.1	--	Posttest only	No
		ACCUPLACER	By class	Posttest (High, Site F)	.79	Large	Posttest only	No
		WorkKeys, Overall	By class	Posttest (High)	.06	Small	Posttest only	No
		WorkKeys	By class	Posttest (High, Site A)	-.34	--	Posttest only	No
		WorkKeys	By class	Posttest (High, Site B)	.07	Small	Posttest only	No
		WorkKeys	By class	Posttest (High, Site C)	.39	Small	Posttest only	No
		WorkKeys	By class	Posttest (High, Site D)	.48	Small-medium	Posttest only	No
		WorkKeys	By class	Posttest (High, Site E)	-.25	--	Posttest only	No
		WorkKeys	By class	Posttest (High, Site F)	.13	Small	Posttest only	No
Walsh-Cavazos, 1994 (2)	QED	PSG Achievement Assessment, Overall	By teacher group (treatment vs. comparison)	Posttest (Elementary)	.556633	Medium	Posttest only	No
		PSG Achievement Assessment, Overall	By teacher group (treatment vs. comparison)	Elementary	.255494	Small	Pretest-Posttest Gain	No

Appendix C: Computation of Effect Sizes, Homogeneity Tests and Q Statistic Analysis

Several computations were carried out to produce effects sizes. For those that were computed using standard formulas, means, standard deviations and sample sizes were entered into pre-set cells in a coding form that calculated effect sizes for **continuous outcome measures using Cohen's *d***

$$d = \frac{(\overline{Y}^{trt} - \overline{Y}^{ctrl})}{S_{pool}},$$

where $\overline{Y}^{trt}$ and $\overline{Y}^{ctrl}$ represent the mean values for the treatment and control groups. S_{pool} was computed as

$$S_{pool} = \sqrt{[(n_i^{trt} - 1)(s_{Yi}^{trt})^2 + (n_i^{ctrl} - 1)(s_{Yi}^{ctrl})^2] / (n^{ctrl} + n^{trt} - 2)},$$

where n_i^{trt} and n_i^{ctrl} are sample sizes and s_{Yi}^{trt} and s_{Yi}^{ctrl} are the standard deviations in study *i*.

The **odds-ratio formula for dichotomous outcome measures**:

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1/q_1}{p_2/q_2} = \frac{p_1 q_2}{p_2 q_1},$$

where p_1 (is the proportion of cases with the outcomes of interest in the first group) and p_2 (proportion in the second group) and $q_x = 1 - p_x$. An odds ratio of 1 shows that the outcome (e.g., achieving math proficiency) under study is equally likely in both groups.

Moreover, effect sizes were computed according to whether the study involved pretest-posttest comparison or only reported posttest results. For **posttest only analysis**, the effect size was computed as the standardized difference between means of the treatment group and the control group on the post means. Specifically,

$$d_{post} = \frac{(\overline{Y}^{trt} - \overline{Y}^{ctrl})}{S_{pool}},$$

where $\overline{Y}^{trt}$ and $\overline{Y}^{ctrl}$ represent the mean posttest values for the treatment and control groups. S_{pool} was computed as

$$S_{pool} = \sqrt{[(n_i^{trt} - 1)(s_{Yi}^{trt})^2 + (n_i^{ctrl} - 1)(s_{Yi}^{ctrl})^2] / (n^{ctrl} + n^{trt} - 2)},$$

where n_i^{trt} and n_i^{ctrl} are the sample sizes for treatment and control group respectively in study *i*, and s_{Yi}^{trt} and s_{Yi}^{ctrl} are the posttest standard deviations for study *i*.

For **pretest-versus-posttest analysis**, the following formula was used to allow for an overall comparison between treatment and control groups, while controlling for the effects of the pretest. Specifically,

$$d_{pre\_post} = \frac{(\bar{Y}^{trt} - \bar{Y}^{ctrl}) - (\bar{X}^{trt} - \bar{X}^{ctrl})}{S_{pool}},$$

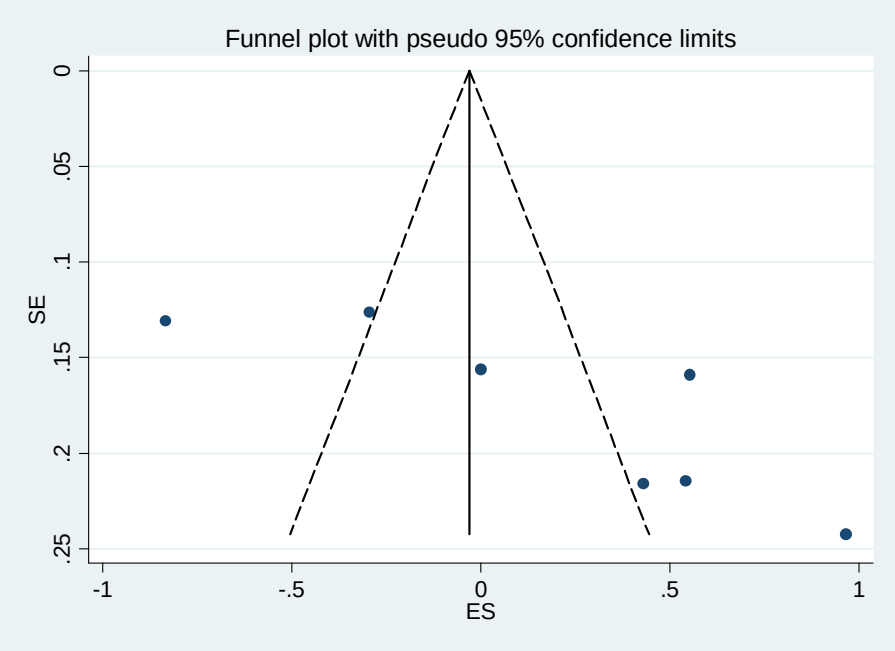
where $\bar{Y}^{trt}$ and $\bar{Y}^{ctrl}$ represent the mean posttest values for the treatment and control groups, respectively, and $\bar{X}^{trt}$ and $\bar{X}^{ctrl}$ represent the mean pretest values for the treatment and control group. S_{pool} was computed as

$$S_{pool} = \sqrt{[(n_i^{trt} - 1)(s_{Yi}^{trt})^2 + (n_i^{ctrl} - 1)(s_{Yi}^{ctrl})^2] / (n^{ctrl} + n^{trt} - 2)},$$

where n_i^{trt} and n_i^{ctrl} are the sample sizes for treatment and control group respectively in study *i*, and s_{Yi}^{trt} and s_{Yi}^{ctrl} are the posttest standard deviations for study *i*.

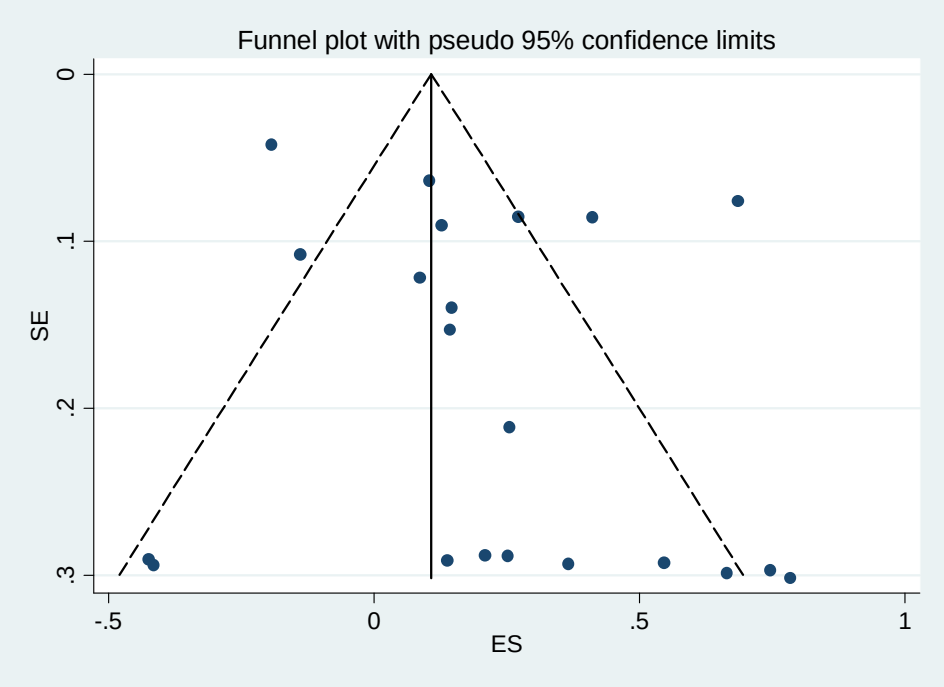
For studies reporting multilevel analyses, effects were computed following Hedges' suggestions when the interclass correlation was reported in the studies (Hedges, 2007).

Homogeneity tests were conducted for each type of measure and subject (math posttest only, math pretest-posttest gains, science posttest only, and science pretest-posttest gains) to determine whether effects from the studied populations are similar or homogeneous. In the case of this meta analysis, the null hypothesis asserts that the effects represent the same population. In all four cases, the null hypothesis is rejected.



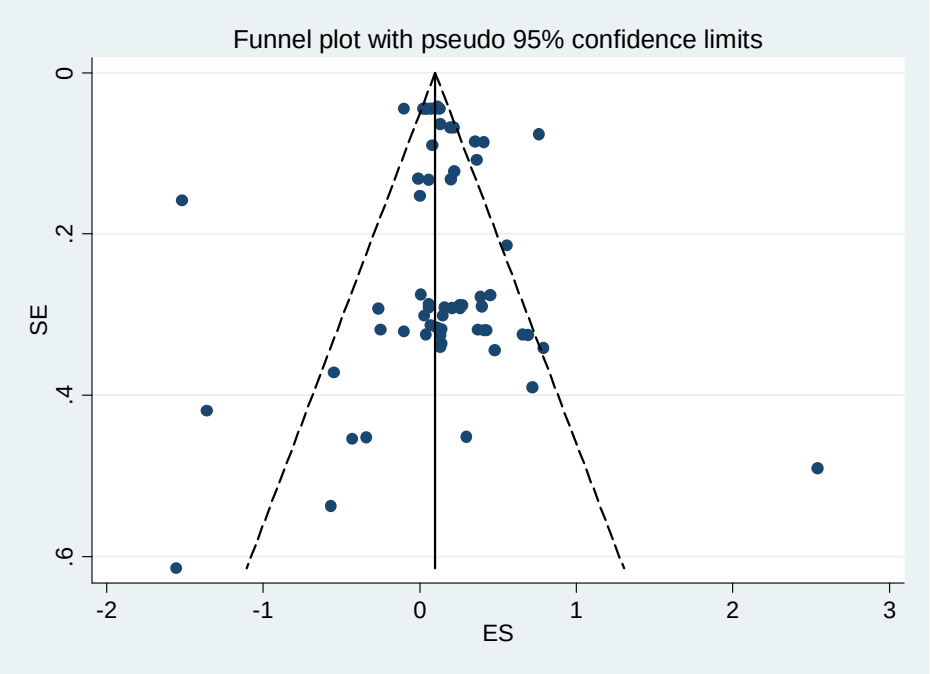
The ***Q*-statistic or *Q* test** was used to assess whether there is true heterogeneity (between-studies variability) in a meta analysis, which in turn affects the statistical model (fixed effects model or random-effects model) used on the meta analysis data to calculate a mean effect size. For this meta analysis, if the studies' results are different due to sampling error, then a fixed-effects model is applied. If the studies' results are different by more than sampling error (considered a heterogeneous case), then a random-effects model was applied. Significance is determined at the p<.05 level. For more information on the *Q* statistic, see Lipsey & Wilson, 2001.

The homogeneity test for the 21 effects from the **math pretest-posttest gains** data set was statistically significant ($Q(20) = 153.71$, $p < .0005$, $I^2 = .870$) indicating that the effects do not represent the same population. The weighted mean effect under the random-effects model for these 21 data points is .210 ($SE = .078$), which indicated that the mean effect differed from zero ($z = 2.70$, $p = .007$) with a 95% confidence interval (*CI*) from .048 to .373. The funnel plot below illustrates the distribution of the effects, and also provides a way to gauge the presence of publication bias. There is some asymmetry, and points appear to be missing in the negative range, suggesting possible bias.

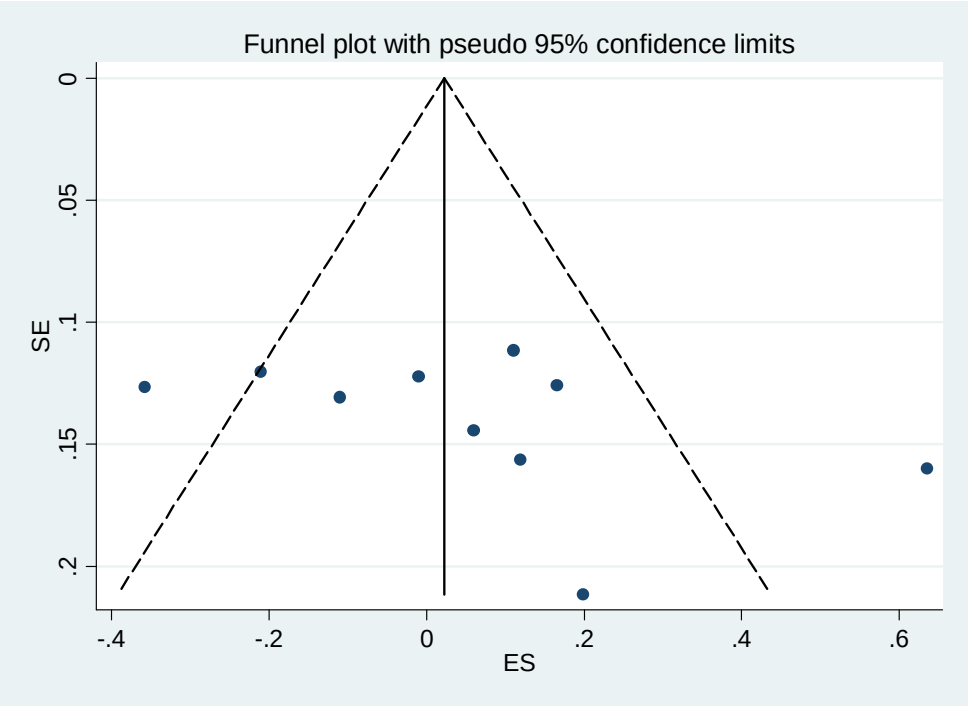


The homogeneity test for the 68 effects from the **math posttest only** data set was statistically significant ($Q(67) = 328.785$, $p < .0005$, $I^2 = .796$) indicating that the effects also do not represent the same population. The weighted mean effect under the random-effects model for these 68 data points is .132 ($SE = .0455$), which indicated that the mean effect differed from zero ($z = 4.05$, $p < .001$) with a 95% confidence interval (*CI*) from .041 to .223. The funnel plot below represents these findings. This plot is less suggestive of publication bias.

Appendix C continued



The homogeneity test for the 10 effects from the **science pretest-posttest gain** data set was statistically significant ($Q(9) = 31.57$, $p < .0005$, $I^2 = .715$) indicating that the effects also do not represent the same population. The weighted mean effect under the random-effects model for these 10 data points is .046 (*SE* = .0838), with a 95% confidence interval (*CI*) from -.143 to .236. Here the mean effect does not differ from zero. The funnel plot below shows these findings, and is too sparse to provide a good assessment of the likelihood of publication bias.



Appendix C continued

The homogeneity test for the 7 effects from the **science posttest only** data set was statistically significant ($Q(6) = 84.15$, $p < .0005$, $I^2 = .929$) indicating that the effects do not represent the same population. The weighted mean effect under the random-effects model for these 7 data points is .176 (*SE* = .237), with a 95% confidence interval (*CI*) from -.404 to .757. Again, this mean does not differ from zero. The funnel plot below shows the distribution of the effects, and the small number of effects precludes making a good assessment of bias.

Appendix D: Correlation Table of Math Post-Only Professional Development Design Elements

	1	2	3	4	5	6	7	8	9	10	11	12	13
Time													
1. Contact Hr.	1												
2. Frequency	.741**	1											
3. Duration	.834**	.623**	1										
PD Activities													
4. Summer Institutes	.577**	.399**	.655**	1									
5. College Courses	.744**	-.171	.596**	.618**	1								
6.Conferences	-.196	.094	.146	-.403**	-.249*	1							
7. Study Group	-.694**	-.253	-.602**	-.524**	-.369**	.287*	1						
Active Learning													
8. Lead Discussion	-.196	.094	.146	-.403**	-.249*	1.000**	.287*	1					
9. Learning Network	-.657**	.048	-.601**	-.351**	-.471**	.249*	.796**	.249*	1				
10.Develop Assessments	-.138	.398**	.135	.345**	-.249*	-.172	.021	-.172	.155	1			
11. Observe Teachers	-.154	.562*	.084	.418**	-.360**	-.249*	-.298*	-.249*	-.093	.692**	1		
12. Classroom Mentoring	-.421**	-.571**	-.742**	-.394**	-.028	-.347**	.579**	-.347**	.502**	-.347**	-.502**	1	
Coherence – 2 Types	.043	-.161	.106	-.406**	-.244*	.221	.163	.221	-.158	-.080	-.324**	-.059	1

In two-tail test: * significant at p<.05; ** significant at p<.01

References

References marked with an asterisk (*) indicate studies included in the meta analysis.

Ball, D. L., & Bass, H. (2000). Interweaving content and pedagogy in teaching and learning to teach: Knowing and using mathematics. In J. Boaler (ed.) *Multiple perspectives on the teaching and learning of mathematics*. (pp. 83–104). Westport, CT: Ablex.

Banilower, E. R., Boyd, S. E., Pasley, J. D., & Weiss, I. R. (2006, February). *Lessons from a decade of mathematics and science reform: A capstone report for the Local Systemic Change through Teacher Enhancement Initiative.* Chapel Hill, NC: Horizon Research, Inc. Retrieved March 21, 2006, from http://www.pdmathsci.net/reports/capstone.pdf

Birman, B. F., & Porter, A. C. (2002). Evaluating the effectiveness of education funding streams. *Peabody Journal of Education, 77*(4), 59–85.

Birman, B. F., Le Floch, K. C., Klekotka, A., Ludwig, M., Taylor, J. Walters, J. et al. (2007). *State and local implementation of the No Child Left Behind Act: Volume II — teacher quality under NCLB: Interim report.* Washington DC: U.S. Department of Education, Office of Planning, Evaluation and Development, Policy and Program Studies Service.

Blank, R. K., de las Alas, N., & Smith, C. (2007, February*). Analysis of the quality of professional development programs for mathematics and science teachers: Findings from a cross-state study.* Washington, DC: Council of Chief State School Officers. Retrieved February 29, 2008, http://www.ccsso.org/content/pdfs/year%202%20new%20final%20NSF%20Impde%20Fall%2006%20%20Report%20-032307.pdf

Blank, R. K., de las Alas, N., & Smith, C. (2008, February). *Does teacher professional development have effects on teaching and learning? Evaluation findings from programs in 14 states.* Washington, DC: Council of Chief State School Officers. Retrieved, March 17, 2009, http://www.ccsso.org/content/pdfs/cross-state_study_rpt_final.pdf

Borko, H. (2004, November). Professional development and teacher learning: Mapping the terrain. *Educational Researcher, 33*(8), 3-15.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2002, November). *Comprehensive school reform and student achievement. A meta analysis.* (Report. No. 59). Baltimore, MD: Center for Research on the Education on Students Placed At Risk, Johns Hopkins University.

Carey, K. (2004, Winter). The real value of teachers: Using new information about teacher effectiveness to close the achievement gap. *Thinking K-16, 8*(1), 3–41.

*Carpenter, T., Fennema, E., Peterson, P., Chiang, C., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal, 26*(4), 499–531.

Chambers, J. G., Lam, I., & Mahitivanichcha, K. (2008, September*). Examining context and challenges in measuring investment in professional development: a case study of six school districts in the southwest region.* (Issues & Answers Report, REL2008-No. 037). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved January 27, 2009, from http://ies.ed.gov/ncee/edlabs/regions/southwest/pdf/REL_2008037.pdf

Choy, S. P., Chen, X., & Bugarin, R. (2006, January). *Teacher professional development in 1999-2000: What teachers, principals, and district staff report.* (NCES 2006-305). Washington, DC: National Center for Education Statistics.

Clewell, B. C., Cosentino de Cohen, C., Campbell, P. B., Perlman, L., Deterding, N., Manes, S., et al. (2004, December). *Review of evaluation studies of mathematics and science curricula and professional development models.* Report submitted to the GE Foundation. Unpublished manuscript.

Coalition for Evidence-Based Policy. (2003). Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide. [Electronic resource]. Washington, DC: U.S. Dept. of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Cohen, D. K., & Hill, H. C. (1998). *Instructional policy and classroom performance: The mathematics reform in California* (RR-39). Philadelphia, Consortium for Policy Research in Education. Retrieved April 29, 2005, from http://www.cpre.org/Publications/rr39.pdf

Corcoran, T. B. (2007). *Teaching matters: How state and local policymakers can improve the quality of teachers and teaching*. (CPRE Policy Briefs RB-48). Philadelphia, PA: Consortium for Policy Research in Education, University of Pennsylvania.

Corcoran, T., & Foley, E. (2003). *The promise and challenge of evaluating systemic reform in an urban district. Research perspectives on school reform: Lessons from the Annenberg Challenge.* Providence, RI: Annenberg Institute at Brown University.

Correnti, R. (2007). An empirical investigation of professional development effects on literacy instruction using daily logs. *Educational Evaluation and Policy Analysis, 29*(4), 262–295.

Council of Chief State School Officers. (2006). *Improving evaluation of teacher professional development in math and science, year 1 project report.* Washington, DC: Author.

Darling-Hammond, L. (1999). *Teacher quality and student achievement: A review of state policy evidence.* University of Washington: Center for the Study of Teaching and Policy. Retrieved April 29, 2005, from http://depts.washington.edu/ctpmail/PDFs/LDH_1999.pdf

Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the United States and abroad.* Washington, DC: National Staff Development Council.

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis, 24*(2), 81–112.

*Dickson, T. K. (2002). *Assessing the effect of inquiry-based professional development on science achievement tests scores.* (Doctoral Dissertation, University of North Texas, 2002). (UMI No. 3076239).

Frechtling, J (2001). What evaluation tells us about professional development programs in math and science. In C. R. Nesbit, J. D, Wallace, D. K. Pugalee, A.-C. Miller, & W. J. DiBiase (Eds.), *Developing Teacher Leaders: Professional Development in Science and Mathematics* (pp. 17–42). Columbus, OH: ERIC Clearinghouse for Science Mathematics, and Environmental Education.

Garet, M. S., Birman, B. F., Porter, A. C., Desimone, L., Herman, R. & Yoon, K. S. (1999). *Designing effective professional development: Lessons from the Eisenhower program and technical appendices* (Report No. ED/OUS99-3). Washington, DC: American Institutes for Research. (ERIC Document Reproduction Service No. ED442634)

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915–945.

Grant, S. G.. Peterson, P. L., & Shojgreen-Downer, A. (1996, Summer). Learning to teach mathematics in the context of systemic reform. *American Educational Research Journal, 33*(2), 509-41.

Guskey, T. R. (2003, June). What makes professional development effective? *Phi Delta Kappan, 84*(10), 748–750.

Harris, D. N., & Sass, T. R. (2007, March). *Teacher training, teacher quality and student achievement.* (Working Paper 3). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research. Retrieved April 12, 2007, from http://www.caldercenter.org/PDF/1001059_Teacher_Training.pdf

Hedges, L. V. (2007, December 1). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics, 32*(4), 341–370.

*Heller, J. I., Curtis, D. A., Rabe-Hesketh, S., Clarke, C., & Verbencoeur, C. J. (2007, August 29). *The effects of "Math Pathways and Pitfalls" on students' mathematics achievement: National Science Foundation final report.* (ERIC Document Reproduction Service No. ED498258). Retrieved November, 9, 2007, from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/31/9b/53.pdf

Hezel Associates, LLC. (2007). *PBS TeacherLine national survey of teacher professional development 2005-2006.* Syracuse, NY: Author.

Hiebert, J. (1999, January). Relationships between research and the NCTM standards. *Journal for Research in Mathematics Education, 30*(1), 3–19.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004, September). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal, 105*(1), 11.

Ingvarson, L., Meiers, M. & Beavis, A. (2005, January 29). Factors affecting the impact of professional development programs on teachers' knowledge, practice, student outcomes & efficacy. *Education Policy Analysis Archives, 13*(10). Retrieved April 29, 2005, from http://epaa.asu.edu/epaa/v13n10/

*Jagielski, D. A. (1991). An analysis of student achievement in mathematics as a result of direct and indirect staff development efforts focused on the problem-solving standard of the National Council of Teachers of Mathematics. (Doctoral Dissertation, Loyola University of Chicago, 1991). (UMI No. 9119821)

Kennedy, M. (1998). *Form and substance in inservice teacher education.* [Research Monograph No. 13]. Madison, WI: University of Wisconsin, Staff National Institute for Science Education.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta analysis*. Applied Social Research Methods Series (Vol. 49). Thousand Oaks, CA: Sage.

Loucks-Horsley, S., Hewson, P., Love, N., & Stiles, K. E. (1998). *Designing professional development for teachers of science and mathematics.* Thousand Oaks, CA: Corwin Press.

McMillan, J. H., & Schumacher, S. (1997). *Research in education. A conceptual introduction.* (4th ed.). New York: Addison-Wesley Educational Publishers, Inc.

*META Associates. (2006, March). *Northeast Front Range math/science partnership (MSP) to increase teacher competence in content. Year 2 evaluation report: January 1, 2005–December 31, 2005.* Golden, CO: Author.

*META Associates. (2007, March). *Northeast Front Range math/science partnership (MSP) to increase teacher competence in content. Final evaluation report: January 1, 2004–December 31, 2006.* Golden, CO: Author.

*Meyer, S. J., & Sutton, J. T. (2006, October). *Linking teacher characteristics to student mathematics outcomes: Preliminary evidence of impact on teachers and students after participation in the first year of the Math in the Middle Institute Partnership.* Paper presented at the MSP Evaluation Summit II, Minneapolis, MN.

Miles, K. H., Odden, A., Fermanich, M., Archibald, S. (2004). Inside the black box of school district spending on professional development: lessons from five Urban Districts. *Journal of Education Finance, 30*(1), 1–26.

National Center on Education Statistics. (n.d.). *Statewide longitudinal data systems grant program: Grantee states.* [Website]. Retrieved March 18, 2009 from http://nces.ed.gov/Programs/SLDS/stateinfo.asp

National Commission on Teaching & America's Future (1996). *What matters most: Teaching for America's future.* New York: Author.

*Niess, M. L. (2005). *Oregon ESEA Title IIB MSP: Central Oregon consortium. Report to the U.S. Department of Education, Mathematics and Science Partnerships*. Corvallis, OR: Department of Science & Mathematics Education, Oregon State University.

Noyce, P. (2006, September 13). Professional development: How de we know if it works? *Education Week, 26*(3), 36–37, 44.

Obama, B. (2009, March 10). Taking on education. Remarks made at the U.S. Hispanic Chamber of Commerce, Washington, DC. Retrieved March 20, 2009, http://www.whitehouse.gov/blog/09/03/10/Taking-on-Education/

O'Reilly, F. E., & Weiss, C. H. (2006, April). *Opening the black box: Using theory-based evaluation to understand professional development for k-12 teachers of math and science*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

*Palmer, E. A., & Nelson, R. W. (2006, September). *Researchers in every classroom. Evaluation report, 2005-06.* Barnes, WI: ASPEN Associates.

*Rubin, R. L., & Norman, J. T. (1992). Systematic modeling versus the learning cycle: Comparative effects of integrated science process skill achievement. *Journal of Research in Science Teaching,* 29, 715–727.

*Saxe, G. B., Gearhart, M, & Nasir, N. S. (2001). Enhancing students' understanding of mathematics: A study of three contrasting approaches to professional support. Journal of *Mathematics Teacher Education, 4*, 55–79.

Scher, L. S., & O'Reilly, F. E. (2007, March). *Understanding professional development for k-12 teachers of math and science: A meta-analysis.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

*Scott, L. M. (2005). The effects of science teacher professional development on achievement of third-grade students in an urban school district. *Dissertation Abstracts International, 66*(04), 1268A. (UMI No. 3171980)

Shavelson, R. J., & Towne, L. (2002). *Scientific research in education.* Washington, DC: National Academies of Sciences. Retrieved April 29, 2005, from http://www.nap.edu/books/0309082919/html/index.html

Showers, B., Joyce, B. & Bennett, B. (1987). Synthesis of research on staff development: A framework for future study and state-of-the-art analysis. *Education Leadership, 45*(3), 77-87.

*Siegle, D., & McCoach, D. (2007). Increasing student mathematics self-efficacy through teacher training. *The Journal of Secondary Gifted Education,* 18(2), 278–331.

*Snippe, J. (1992, July). *Effects of instructional supervision on pupils' achievement.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. Unpublished manuscript.

Supovitz, J. A. (2003). Evidence of the influence of the National Science Education Standards on the professional development system. In K. S. Hollweg & D. Hill (Eds.), *What is the influence of the National Science Standards?* (pp. 64–75). Washington, DC: National Academy Press.

*Walsh-Cavazos, S. (1994). A study of the effects of a mathematics staff development module on teachers' and students' achievement. *Dissertation Abstracts International, 56*(01), 165A. (UMI No. 9517241)

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008, November). Experimenting with teacher professional development: Motives and methods. *Educational Researcher, 37*(8), 469–479.

Weiss, I. R., Banilower, E. R., McMahon, K. C., & Smith, P. S. (2001). *Report of the 2000 national survey of science and mathematics education.* Chapel Hill, NC: Horizon Research, Inc. Retrieved April 29, 2005, from http://2000survey.horizon-research.com/reports/status/complete.pdf

Wilson, S. M., & Berne, J. (1999). Teacher learning and the acquisition of professional knowledge: An examination of research on contemporary professional development. *Review of Research in Education, 24*, 173–209.

Yoon, K. S., Duncan, T., Lee, S., W.-Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement.* [Issues & Answers Report, REL 2007-No. 033]. Washington, DC: U.S. Department of Education, Institute of Educational Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved November 1, 2007 from http://ies.ed.gov/ncee/edlabs/regions/southwest/pdf/REL_2007033.pdf