CRESST REPORT 795

Eva L. Baker

PROGRESS REPORT YEAR 4: NATIONAL CENTER FOR RESEARCH ON EVALUATION, STANDARDS, AND STUDENT TESTING (CRESST) THE DEVELOPMENT AND IMPACT OF POWERSOURCE©

MAY, 2011

The National Center for Research on Evaluation, Standards, and Student Testing

Graduate School of Education & Information Sciences
UCLA | University of California, Los Angeles

Progress Report Year 4:
National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
The Development and Impact of POWERSOURCE©

CRESST Report 795

Eva L. Baker, Principal Investigator
CRESST/University of California, Los Angeles

May, 2011

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

Copyright © 2011 The Regents of the University of California.

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number Award Number R305A050004, as administered by the Institute of Education Sciences, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences, or the U.S. Department of Education.

To cite from this report, please use the following as your APA reference:
Baker, E. L. (2011). *Progress Report Year 4: National Center for Research on Evaluation, Standards, and Student Testing (CRESST) The Development and Impact of POWERSOURCE©.* (CRESST Report 795). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

TABLE OF CONTENTS

PROGRESS REPORT YEAR 4:

NATIONAL CENTER FOR RESEARCH ON EVALUATION, STANDARDS, AND STUDENT TESTING (CRESST)

THE DEVELOPMENT AND IMPACT OF POWERSOURCE©

Eva L. Baker, Principal Investigator
CRESST/University of California, Los Angeles

Description of POWERSOURCE© and its Rationale

The POWERSOURCE© intervention is intended as a generalizable and powerful formative assessment strategy that can be integrated with any on-going mathematics curriculum to improve teachers' knowledge and practice and, in turn, student learning. Combining theory and research in cognition, assessment and learning (for both adults and students) with design elements to support the transformation of practice within existing constraints, POWERSOURCE© includes both a system of learning-based assessments and an infrastructure to support teachers' use of those assessments to improve student learning. The current study focuses on middle school mathematics, starting in grade 6, and on helping to assure that students possess key understandings they need for success in Algebra I. Such a focus is motivated by ample research showing the frequency and price of failure for subsequent academic performance, including high school graduation, college entry and preparation (e.g., Brown & Niemi, 2007). Our primary research objectives are based on our hypotheses that as a result of POWERSOURCE©, teachers will become more proficient in their subject matter knowledge, more skilled in their formative use of assessment, and better focus their instruction on key ideas, and, as a result, will be more effective in helping students to improve their understanding, as shown by measures of student learning. Ultimately, we expect the improvements in student understanding to drive better performance on No Child Left Behind (NCLB) mandated state tests, transfer measures, and future coursework.

Research on Formative Assessment

The intervention builds on recent research showing formative assessment as a powerful strategy for improving learning. (Black & Wiliam, 1998a, 1998b; Bloom, 1968; Kluger & DeNisi, 1996). For example, Black and Wiliam's (1998a) landmark meta-analysis, based on a review of 250 studies, found effect sizes that ranged between .4 and .7, and found particularly large effect sizes for low-achieving students, including students with learning disabilities (Black & Wiliam, 1998b). This finding makes intuitive sense, as one of the major functions of formative assessment is to determine where students are relative to learning goals and to use this information to provide feedback and/or make necessary instructional adjustments, such as re-teaching, trying alternative instructional approaches, or offering more opportunities for practice. If students have already mastered the content, there is little need for subsequent adjustment and little room for learning improvement.

Yet even as research shows the rich potential of formative assessment, so too it suggests the limits of current practice. The quality of increasingly popular interim or benchmark testing, marketed as formative assessments to districts and schools, is uneven, and assessment tends to be an afterthought rather than a core, quality element of current curriculum materials (Herman & Baker, 2006; Herman, Osmundson, Ayala, Schneider, & Timms, 2006; Wolf, Bixby, Glenn, & Gardner, 1991). Moreover, educators often have limited background and capacity to develop or engage in quality assessment practices (Heritage & Yeagley, 2005; Herman & Gribbons, 2001; Plake & Impara, 1997; Shepard, 2001: Stiggins, 2005). For many teachers, their current classroom assessment practices are almost exclusively summative, consisting, for example, of end-of-the-week, unit or semester tests.

Students receive grades or scores on these assessments and their teachers—who have neither the time nor the curriculum resources to remediate deficiencies--move on, disconnecting the assessments from any active function in learning. Yet as Black and Wiliam (1998a, 1998b) note, assessments can only become formative when information from them is used immediately to inform teaching and for the benefit of student learning. Teacher subject matter knowledge offers yet another challenge, as research and our own experiences in assessment development with teacher and districts suggest that many teachers do not have subject area knowledge sufficiently deep to teach or assess mathematics effectively (Ball & Bass, 2001; Ball, Lubienski, & Mewborn, 2001).

Learning to use assessment in a more formative way thus requires significant changes for many districts, teachers and students. For districts it will mean insuring that teachers have the time and resources to act on the assessment information they receive. For teachers and students, it will involve learning to use assessment information diagnostically to determine the course of instruction and learning, and to deal with learning difficulties that are revealed by formative assessments. Given the challenges involved in changing assessment practices, a substantial part of our research and development thus focuses on exploring the types and frequency of assessments and instructional supports that will be feasible to implement and most beneficial to teachers and students, helping teachers, for example, to understand mathematical concepts more deeply, monitor learning of key ideas and skills, and to figure out the best strategies to improve students' understanding.

Learning-Based POWERSOURCE© Strategy

The POWERSOURCE© intervention thus involves not only the development of formative assessments, but also the development of professional development and instructional support resources to help teachers to understand the mathematical content, interpret assessment information, provide feedback to students, and adapt instruction as needed. Moreover, a striking innovation in POWERSOURCE© is its targeting of the big ideas – fundamental concepts and principles – and their interrelationships that underlie and define a field of knowledge, rather treating specific concepts and topics in isolation, as do traditionally developed tests. This innovation is motivated by ample evidence from a range of cognitive psychology perspectives which suggest that for learning to be acquired efficiently and sustained, it must enable students to connect to organizing principles what otherwise would be disconnected knowledge or procedures and to integrate and demonstrate their knowledge and skills in many situations, in near and far transfer, and across time (e.g., Atkinson & Shiffrin, 1968; Chi, Feltovich, & Glaser, 1981; Ericson, 2003; Ericson & Simon, 1984; Hiebert & Carpenter, 1992; Mayer, 2003; Brown, Bransford, & Cocking, 2000; Newell, 1990, VanLehn, 1996, Catrambone & Holyoak, 1989).

Similarly, the specific item types used in POWERSOURCE© were developed based on cognitive research demonstrating the value of specific strategies for promoting transfer. Research, for example, suggests that learning and problem solving strategies can be successfully transferred if students are taught to focus on: self-evaluation or metacognition (Moreno & Mayer, 2005; Palincsar & Brown, 1984; Pressley & Brainerd, 1985); the conditions for applying strategies (Judd, 1908, 1936; Kilpatrick, 1992), building principled representations of problem situations (Fuchs, Fuchs, Finelli, Courey, & Hamlett, 2004; Kilpatrick, Swafford, & Findell, 2001); use worked-out examples as a way to build problem schemas that generalize across a range of tasks (Chi & Bassok, 1989; Pawley, Ayres, Cooper, & Sweller, 2005); explanation and problem solving tasks requiring understanding of core concepts and principles that recur across arithmetic, pre-algebra, and algebra (Carpenter & Franke, 2001; Haverty, 1999; Ready, Edley, & Snow, 2002; Schmidt, McKnight, & Raizen, 1997). POWERSOURCE© not only uses item types that are positioned to uniquely foster learning, but it also purposively employs multiple formats to promote transfer, rather than focusing only on those representations adopted by test developers designing for accountability purposes (Richardson-Klavehn & Bjork, 2002).

Targeted Domains Operationalized in *Checks for Understanding*

The POWERSOURCE© intervention targets big ideas and related skills in four domains underlying success in Algebra 1: a) rational number equivalence (RNE), b) properties of arithmetic (PA; the distributive property), c) principles for solving linear equations (SE); and d) application of core principles in these domains to other critical areas of mathematics, such as geometry and probability (RA). These domains were chosen because of their importance to later mastery of algebra and their significant place in state mathematics standards across grades 6-8.

In each domain we have designed a series of short POWERSOURCE© assessments comprised of multiple item types, which are called "*Checks for Understanding*," to help teachers assess their students understanding of basic mathematical principles and to connect their instruction and provide feedback to support deeper understanding. A set of instructional resources and targeted professional development activities were also developed for each of these domains. Thus, a POWERSOURCE© module around a given domain includes a set of *Checks for Understanding*, targeted instructional resources, and professional development opportunities. POWERSOURCE© materials are designed to complement existing curricula, but time for it must be found within tight district curriculum frameworks and timelines. It is thus important for POWERSOURCE© to integrate well and easily with existing initiatives and not add an unreasonable burden to the heavy testing requirements already imposed on teachers (e.g., weeks of state and district testing), and not replace large chunks of extant curricula.

More detailed information about the research foundations, content focus, initial development process, and program components of POWERSOURCE© can be found in the Center's 2006, 2007 and 2008 progress reports to the Institute of Education Sciences (Baker, 2006, 2007, 2008). The present report focuses on providing an update on project activities undertaken since the last progress reporting period (i.e., covering the 2008-09 school year). This update is organized around four general areas:

1. First, we provide updated results from the 2006-07 experimental (randomized) field test of POWERSOURCE© instructional sensitivity, including both item quality data for the *Checks for Understanding* and treatment/control differences on student and teacher outcomes.
2. Second, we describe the experimental (randomized) study conducted during the 2007-08 school year, and present findings on both student and teacher outcomes.
3. Third, we describe design and pilot testing of 8th grade materials conducted during the 2008-09 school year.

4. Finally, we provide updates on supplemental/synergistic research studies. Dissemination activities are also discussed, as well as an overview of planned activities for the 2009-10 school year.

Updated Results from 2006-07 POWERSOURCE© Field Test

As described in previous progress reports (Baker, 2007, 2008) during the 2006-07 school year we conducted two types of inter-related studies building on the prior project years' work: a) continuing investigation of item quality of the *Checks for Understanding*, supplementing data we obtained on items in the previous year; and b) experimental tests (using random assignment) of the 6th grade POWERSOURCE© materials in four districts in two states. Detailed description of methodology used and preliminary results were presented in these previous progress reports. Following is an updated summary of 2006-07 school year results from these two inter-related strands of work.

Item Quality Analyses of 2006-07 Field Test Data

Analyses of item and test data from 2006-07 field test continued and were completed over the course of the current year. Several kinds of item-level analyses were carried out: confirmatory factor analyses, reliability analyses, and Item Response Theory (IRT) analyses. Our typical scheme for analyzing each set of *Checks for Understanding* was to first calculate reliability coefficients (Cronbach's alpha) for the items comprising the set. Second, as another check of item quality, we conducted a principal component analysis and a confirmatory factor analysis for each test form to check whether the items exhibited the factor structure we expected; for example, whether the computation items loaded on the same factor. Third, IRT analyses based on Rasch models were conducted in order to obtain item parameters (difficulties) and item characteristic and information curves so that we could use them to select items for future testing. The model-data fit was investigated using two model fit indices. One is the G2 index which is the Chi-square ($\chi 2$) statistic and provided in PARSCALE phase 2 outputs, and the other is the mean square fit (MNSQ) statistics.

Data addressed the question of whether relatively short assessments can provide reliable and useful information on middle school students' understanding of conceptual domains in pre-algebra. Items and test forms were developed and tested in four domains (rational number equivalence, properties of arithmetic, principles for solving equations, and applications of these concepts to other domains), all of which are critical to eventual mastery of algebra. We tested the items with sixth grade students in classrooms in four districts. We then pared down the items to create eight assessment forms that were further tested alongside instructional support materials and professional development. Results of this study suggest that relatively brief formative assessments focused on key conceptual domains can provide reliable and useful information on students' levels of understanding and possible misunderstandings in the domain (see Appendix A for details on this study).

Experimental Comparison Findings

As noted in the earlier text, during the 2006-07 school year we field tested both the POWERSOURCE© assessments and associated instructional materials as part of a random assignment study. Analyses of experimental comparisons for all POWERSOURCE© units were completed during the current year. As noted in the previous annual report, two districts from each Arizona and California were recruited for the study and within these districts, 7 Arizona and 18 California middle schools agreed to participate. Within these schools, 25 6th grade teachers in Arizona and 41 6th grade teachers in California and their students comprised the original study sample. Within each district, teachers were randomly assigned to experimental (POWERSOURCE©) and comparison groups. Experimental group teachers in all cases participated in initial summer professional development and after school follow-up sessions, and used project materials, including the *Checks for Understanding* and instructional supports (including teacher instructional handbooks), but comparison group experiences varied slightly depending on district need and configuration. All teachers gave eight *Checks for Understanding* throughout the school year, two for each of the four POWERSOURCE© modules (RNE, PA, SE, and RA; note that based on district curricula and needs a slightly modified module, FM, was used in the AZ districts in place of the RA module).

Additional details about the 2006-07 field test design, materials, and sample can be found in Baker (2007). Preliminary results from this field test were presented in these progress reports based on the most complete data available at the time. The appended technical report (Appendix B) elaborates on the updated findings from the 2006-2007 field test.

POWERSOURCE© Implementation Study 2007-08

As described in previous reports (Baker, 2007, 2008) the core undertaking of our work during the 2007-08 school year was conducting an extended, random assignment implementation study of our 6th grade POWERSOURCE© program. As with the 2006-07 field test, teachers were randomly assigned to either POWERSOURCE© or control conditions with the ultimate goal of determining program impact on both students and teacher learning outcomes. The 2007-08 study differed from the previous year's work in a number of important ways, however. Specifically:

- The experimental design incorporated both within- and between-school random assignment models. That is, for some of the districts the random assignment accomplished within each school (i.e., a given school had both POWERSOURCE© and control teachers), and for some the random assignment was between school (i.e., all teachers at a given school were POWERSOURCE© or control). Additional background and rationale for this approach can be found in CRESST supplement design report submitted to IES in August, 2007 (Ultimately, a total of 112 6th grade teachers across 7 school districts agreed to participate in the study.
- Although the content focus of the four POWERSOURCE© modules remained the same (RNE, PA, SE, and RA), based on teacher feedback and our implementation experiences in 2006-07 the structure of each unit changed somewhat. In 2007-08, POWERSOURCE© teachers were provided with three *Checks for Understanding* for each unit – one prior to the first day's set of instructional materials, one in between the first and second day of instruction, and one after the second day of instruction. Thus, the students completed 12 *Checks for Understanding* (three for each of the four units) during the school year.
- Unlike 2006-07, the control students did not complete any of the *Checks for Understanding* (i.e., the short formative assessments). Thus, the 2007-08 control students and teachers had no exposure to any of the POWERSOURCE© materials or concepts during the school year.
- All students (POWERSOURCE© and control) completed a test of prerequisite knowledge at the beginning of the school year and a transfer measures of math knowledge at the end of the school year. The test of prerequisite knowledge serves as a baseline measure for later analyses, while the transfer measure will serve as an independent, student outcome measure (in addition to state test data).
- Based on district response and feedback, districts were offered the option of the control teachers receiving an alternative (i.e., non-POWERSOURCE©) professional development from CRESST (as opposed to the control teachers not receiving any additional professional development than what the district already had planned). The majority of participating districts selected this option.

Additional details about the plan and its rationale can be found in the supplemental design report submitted to IES in August, 2007. Appendix C presents the updated results for the 2007-2008 study year.

Measure Quality, Item Analysis, and Test Equating

This section documents the technical characteristics of the reliability, validity, item analysis, item parameter linking and test equating for some of the 2007-2008 POWERSOURCE© *Checks for Understanding* and the pre-test and postttest (transfer measure) administered to all 6th grade students. To examine the measure quality, alpha was used to calculate reliability and exploratory factor analysis was applied to check the construct validity. To investigate the quality of test items, two different angles according to different theories could be applied: One is classical test theory (CTT) and the other is IRT. Because both CTT and IRT can provide valuable information about a test, we used them to evaluate the items in the POWERSOURCE© *Checks for Understanding* assessments (for more information on CTT and IRT see Appendix D).

In keeping with findings in Phelan, Kang, Niemi, Vendlinski, & Choi (2009) which showed the appropriateness in using unidimensional Rasch models for POWERSOURCE© test items, the one parameter logistic model (1PLM) for dichotomous items and partial credit model (PCM; Masters, 1982) for polytomous items were employed. Additionally, with the data sets for the pretest and transfer measure having items representing all domains, factor analysis was conducted to estimate the amount of variance explained by the main construct.

Reliability

Table 1 shows the number of items, the actual number of examinees, and reliability for each form considered in the report. All of the data sets are not yet complete and so included in this analysis are the pretest, transfer measure, and the RA assessments. The reliability was computed with coefficient alpha as shown in Table 1.

Table 1

Sample Size and Reliability of the 2007-2008 POWERSOURCE© Assessments

POWERSOURCE© assessments	# items (= #dichotomous items + #polytomous items)		Sample size	Reliability (Cronbach's alpha)
Pre-test	28	(=28 + 0)	5,838	.80
Transfer measure (Post-test)	31	(=30 + 1)	5,358	.86
RA-ass1	10	(=8 + 2)	3,336	.77
RA-ass2	5	(=4 + 1)	3,272	.57
RA-ass3	5	(=4 + 1)	3,261	.66

Construct Validity

Even though exploratory linear factor analysis (FA) using a product moment correlation matrix is commonly applied to assess construct validity of educational and psychological tests, this method is not appropriate for item-level data with categorical response format because traditional FA assumes continuous ratings and normality. For the purpose of conducting FA for 2007-2008 pre-test and transfer measure data, therefore, full-information item FA was used because it is known to be free from such problems (Bock, Gibbons, & Muraki, 1988).

First of all, for both 2007-2008 pre test and transfer measure data sets, the first factor had much bigger eigenvalues (explaining about 25-30% of total variance) than those of the other factors and was related to all four domain areas such as RA, RNE, PA, and SE. This finding could justify the use of unidimensional IRT to analyze the POSWERSOURCE© test data. Because both pre-test and transfer measure include items from all domain areas of interest, however, it could be possible that considering more than a single construct was required in the use of an IRT model to have better model-fit for the given data. To investigate this possibility, one-, two-, three-, four-, and five-factor models were compared using two model selection indices. One is Akaike's Information Criterion (AIC; Akaike, 1974), and the other is Schwarz's Bayesian Information Criterion (BIC; Schwarz, 1978). Also, a Chi-square test was applied. Tables 2 and 3 contain the comparison results for the 2007-2008 pre-test and transfer measure, respectively.

Table 2

Comparison of Various Factor Models with 2007-2008 Pre-test Data

Factor model	CHI-SQUARE	DF	CHI-Square Difference	DF difference	AIC	BIC
1	56175.17	5781			56287.17	56386.08
2	56316.20	5754	-141.03	27	56482.20	56628.80
3	58205.89	5728	-1889.69	26	58423.89	58616.41
4	55401.62	5703	2804.27	25	55669.62	55906.30
5	56099.09	5679	-697.47	24	56415.09	56694.16

Table 3

Comparison of Various Factor Models with 2007-2008 Transfer Measure Data

Factor model	CHI-SQUARE	DF	CHI-Square difference	DF Difference	AIC	BIC
1	76789.34	5295			76913.34	77020.538
2	74873.00	5265	1916.34	30	75057.00	75216.068
3	74289.31	5236	583.69	29	74531.31	74740.519
4	74057.74	5208	231.57	28	74355.74	74613.361
5	74362.71	5181	-304.97	27	74714.71	75019.014

Among the five factor models (1 through 5), in both Tables 2 and 3, the four factor model was chosen as the best model because this model had the smallest AIC and BIC values. Also, the model with five factors did not show significant improvement over the four factor model according to Chi-square tests.

For the pre-test data, under the four factor model, 37.32% of total variance was explained by the four factors. Also, 24.37%, 5.65%, 4.35%, and 2.95% of total variance was explained by the first, second, third, and fourth factors, respectively. The second factor mainly had large loading values for RNE and SE, except the item, PRE22. The third and fourth factor seemed to mostly measure PA and RA, respectively. The first factor has relatively large correlation (larger than .45) with the three other factors. And the third and fourth factors had relatively small correlation (.192).

For the transfer measure data, under the four factor model, 38.97% of total variance was explained by the four factors. And, 29.39%, 4.45%, 3.31%, and 1.82% of variance was explained by the first, second, third, and fourth factors, respectively. The first factor had much bigger eigenvalue than those of the other factors and was primarily related to RNE, SE, and "others" domains except item POST22. The second factor was clearly about item 29, and the third and fourth factors were mostly related to PA and SE, respectively. The first factor had relatively large correlation with the second and third factors, and the second and fourth factors appeared to have small correlation (.146). With the results found in the earlier text, even though the use of a unidimensional IRT model could be justified, it appeared to be possible that the application of more complicated psychometric models might improve the model-data fit.

Item Analysis

Classical test theory. A descriptive analysis was used initially which contained mean and standard deviation of the test score. Each item was examined using the proportion which answered the item correctly, *p*-values, and point-biserial correlation, r_{pbis}. The former and latter provide the information of item related to difficulty and discrimination, respectively. The point-biserial correlation is the correlation between the test-takers’ performance on one item compared to the test-takers’ performances on the total test score.

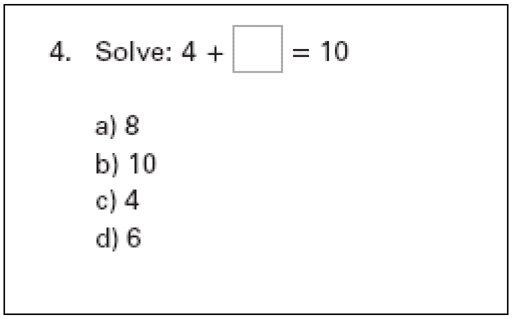
Item response theory. Because the 1PLM having every item discrimination to be 1 is nested within the PCM, the IRT model used in this report can be written:

$$P(z \mid \theta_j, \beta_i, \tau_{ci}) = \frac{\exp \sum_{c=0}^{z} \left[\theta_j - (\beta_i - \tau_{ci}) \right]}{\sum_{y=0}^{Z_i} \exp \sum_{c=0}^{y} \left[\theta_j - (\beta_i - \tau_{ci}) \right]} \tag{1}$$

Under the PCM, the probability that an examinee *j* scores *z* with *z* = 0, …, Z_i on item *i* with $Z_i + 1$ response categories. β_i denotes the difficulty of item *i*, and τ_{ci} represents the location parameter for a category on item *i*. Equation 7 needs to set $\tau_{0i} = 0$, $\sum_{c=1}^{Z_i} \tau_{ci} = 0$ and $\exp \sum_{c=0}^{0} \left[\theta_j - (\beta_i - \tau_{ci}) \right] = 1$ for model identification. For a dichotomous item with $Z_i = 1$, there are two response categories (i.e., 0 and 1) and only β_i exists as the related item parameter.

2007-2008 POWERSOURCE© Pre-test.

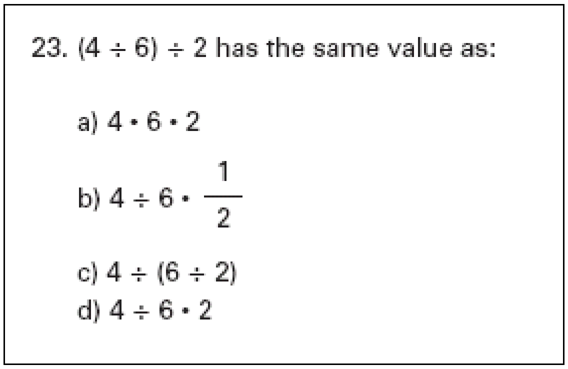
Among the 28 items on the 2007- 2008 POWERSOURCE© pre-test, the item PRE04 was the easiest item (b= -2.600, p-value=0.99), and PRE23 was the most difficult item (b=3.139, p-value=0.13; see Figure 1 for these items).



4. Solve: 4 + □ = 10

a) 8
b) 10
c) 4
d) 6

Item Pre-04



23. (4 ÷ 6) ÷ 2 has the same value as:

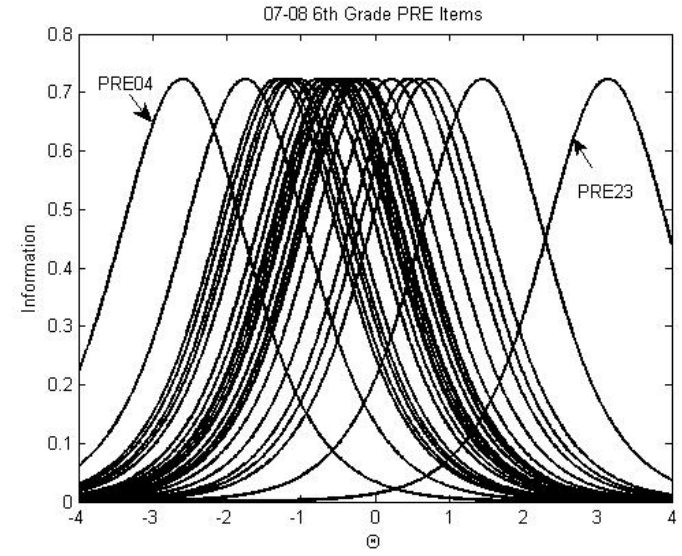
a) 4 • 6 • 2
b) 4 ÷ 6 • $\frac{1}{2}$
c) 4 ÷ (6 ÷ 2)
d) 4 ÷ 6 • 2

Item Pre - 23

Figure 1. Easiest and most difficult items on the Grade 6 Pretest.

The polyserial correlation coefficients between item and test scores were larger than 0.3 except for two items (PRE23 and PRE24). These two appeared to have poor discrimination, and were deemed poor quality items. The test and item reliability based on item response theory were calculated. The IRT test reliability was calculated as Dimitrov (2003) suggested and it was .917 (Cronbach's alpha = .80). And, the most difficult item (PRE23) had the smallest item reliability, which means it has less contribution to test reliability than the other items in the pretest. See Appendix E for the full results of the Item Analysis of the POWERSOURCE© Pretest and see Appendix F for the complete set of pretest items used.

The item information curves in Figure 2 show that PRE04 and PRE08 mainly give information for the examinees with low ability. And the difficult items PRE27 and PRE23 are providing relatively large amounts of information for the examinees with high ability.



From left to right : PRE04 PRE08 PRE03 PRE07 PRE09 PRE13 PRE18 PRE25 PRE19 PRE12 PRE10 PRE14 PRE01 PRE22 PRE28 PRE06 PRE16 PRE17 PRE02 PRE11 PRE15 PRE20 PRE26 PRE21 PRE05 PRE24 PRE27 PRE23

Figure 2. The item characteristic curves of POWERSOURCE© pretest items.

Implementation study 2007-08: transfer measure. In 2007-2008 5,358 students completed the transfer measure. Item analyses were carried out on 29 transfer measure items (there were 31 possible responses as 1 item had 3 parts). Among the 31 items (30 dichotomous and 1 polytomous), we determined degree of difficulty from easiest (b=-1.629, p-value=0.911) to most difficult (b=1.296, p-value=0.141). The polyserial correlation coefficients between item and test scores were larger than 0.3 except for one item that had poor discrimination. The test and item reliability based on item response theory were also calculated. The test reliability was .93 (Cronbach's alpha = .86). The polytomous item

(POST27) had the largest item reliability, which means it has more contribution to test reliability than the other items in transfer measure. The item information curves also indicated that the explanation task provided the largest amount of information (see Figure 3).

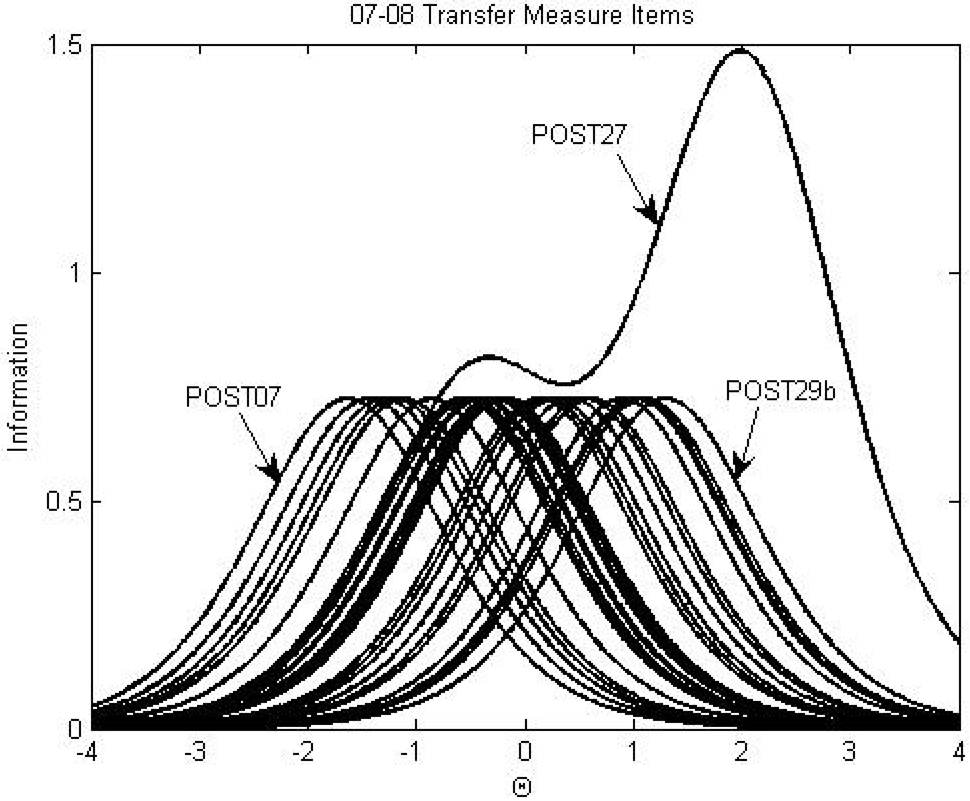


Figure 3. The item characteristic curves of POWERSOURCE© Transfer Measure items.

As shown in the figure, the easiest item (POST07) mainly gives information for the examinees with low ability and the most difficult item (POST29b) is providing more information for the examinees with high ability. See Appendix G for the complete set of transfer measure items used. The results of item analysis for three assessments in review and applications are given in Appendix H.

Item Parameter Estimation, Linking, and Equating

Non-equivalent groups anchor-test design. In the NEAT design, the different forms (i.e., Forms A and B in Figure 4) and are administered to different groups of test takers (i.e., Groups P and Q). And, the groups are not assumed to be equivalent. Even though the same examinees took the various 2007-2008 POWERSOURCE© test forms, they were expected to have changes in their ability due to the effect of instructional intervention. Across the test administrations, therefore, the examinees were assumed to make up different groups with different ability distributions. The common items were required between forms to place item parameters onto the same metric given those group differences. These common item sets (anchors) should be chosen to represent the content and statistical characteristics of the test. The POWERSOURCE© test forms, however, did not have common items. To solve this problem, an external anchor form (EAF) was newly created and administered to a different group of 6th graders (*N*=450 from two districts: one in Nevada and one in California). The EAF contained 32 items and shared at least two common items with every test form in 2007-2008 POWERSOURCE©.

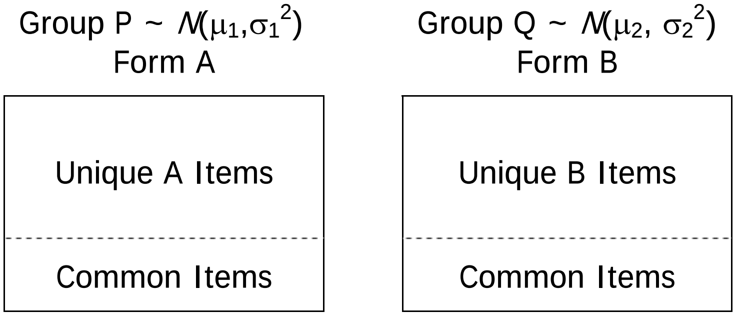


Figure 4. NEAT Data Collection Design for Common Item Equating.

Item parameter estimation and linking. To estimate the parameters, the computer program PARSCALE (Muraki & Bock, 1997) was used. There, the item parameters are estimated with marginal maximum likelihood estimators. Once the item parameters of the EAF were estimated, the next PARSCALE run was conducted for each operational form with the parameters of common items held fixed as obtained from the previous PARSCALE run. The EAF and every other form worked as Form A and Form B in Figure 4, respectively. And, the item parameters of every POWERSOURCE© test form (i.e., Form B) could be put onto the same scale through using the FIPC method (see Appendix I for the item analysis results for the EAF items).

Test equating. The various test-score equating methods under the NEAT design are distinguished in terms of their statistical assumptions (e.g., see Kolen & Brennan, 1995). The IRT true-score equating method (Lord, 1982) was used in equating POWERSOURCE© test forms. Appendix I and Figure 5 show item analysis results and the test characteristic curve (TCC) of the EAF, respectively. The expected true scores (actually, after rounding to the nearest whole number, 33 integer values from 0 to 32) were used as the scale scores of every POWERSOURCE© test form. The raw scores of each POWERSOURCE© form are transformed into the scale scores. In other words, for each form, the equating conversion table was established.

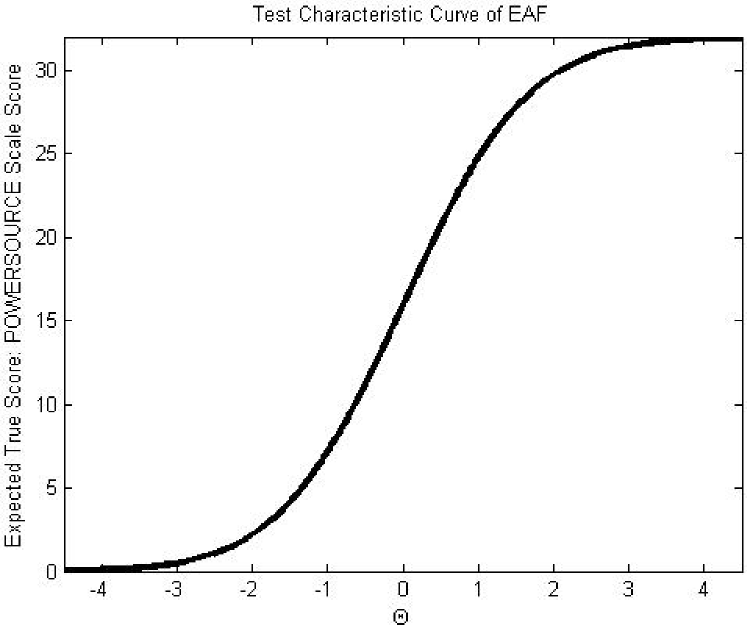


Figure 5. The TCC of EAF as a Basis to build up POWERSOURCE© Scale Scores.

2007-2008 POWERSOURCE© pre-test and transfer measure. The linked item parameters (i.e., placed onto the EAF scale) for these two test forms are shown in Tables 3 and 4, respectively. The items with fixed parameters that came from the PARSCALE run of the EAF data were shaded in the tables. Each test form had four such items and the other item parameters were estimated to be put onto the scale of EAF.

IRT true score equating discovers the equivalent score on the 0 to 32 (EAF) scale score metric, $\varphi(x)$, for an observed score x on a 2007-2008 POWERSOURCE© test form (X) using the test characteristic curves for both forms (EAF and X) which respectively define the relationship between person location parameters (i.e., θ) and the corresponding true test scores. Appendix J provides the one-to-one relationship between raw scores and POWERSOURCE© scale scores. The raw-to-scale conversion tables for the *Checks for Understanding* assessments for the Review and Applications domain provided in Appendix K.

Implementation Study 2007-08: Student Outcomes

As described earlier, we incorporated both within- and between-school random assignment models. These two designs were based on based on district needs and configuration. Ultimately, three of the districts used a within-school (W-S) design, where random assignment was accomplished within each school (i.e., a given school had both treatment and control teachers). Four districts used a between-school (B-S) design, where schools within a district were randomly assigned to treatment or control conditions.

Eighty-five teachers from 25 schools in the 7 school districts participated in the study. All teachers taught 6th grade. Table 5 shows the distribution of teachers in each district.

Table 5

Sample Distribution ('07-'08 school year)

Ct/Tr	*N* of students	*N* of teachers	*N* of schools
Between design			
Control	633	13	8
Treatment	842	23	7
Subtotal	1475	36	15
Within design			
Control	1120	23	5
Treatment	1496	26	5
Subtotal	2616	49	10

Table 6

Sample Distribution by School District ('07-'08 school year)

District	*N* of students	*N* of teachers	*N* of schools	Design
AZ-1	93	2	1	B/S
CA-1	872	18	3	W/S
CA-2	770	11	2	W/S
CA-3	195	5	3	B/S
CA-4	279	11	5	B/S
CA-5	974	20	5	W/S
CA-6	908	18	6	B/S

Due to administrative reasons, many pretest scores in AZ-1 district were not valid and excluded from the analyzed sample as missing values. Accordingly, the number of sample in this district was substantially small compared to the numbers in the other school districts. Note that two of the treatment schools in our initial sample of B-S design were removed from the analysis owing to issues of noncompliance with study procedures.

HLM results. Taking methodological concerns into account, we used a two-level hierarchical model (HM) to examine the POWERSOURCE© effects on the transfer measure outcome. In order to synthesize two different designs and compromise unit of analysis issue, we chose teacher as a unit of analysis and individual school effects are also included in a model. School specific fixed effects take care of school blocking factors and intra-class correlation of school in a model. As such, we can examine whether there is a differential treatment effect depending upon two different designs not at the cost of losing statistical power (see Appendix L for the complete statistical model used).

Results indicate that scores on the pretest were slightly higher in the POWERSOURCE© group than in the control group. Thus, we used these initial differences between the groups as a covariate in our statistical model.

Using the Transfer Measure Total Score as Student Outcome

Figure 6 shows the relationship between student pretest and transfer measure for POWERSOURCE© vs. Control, and within and between school design. In the between-school design, the two fitted line cross approximately -1.0 SD of pretest score, while in the within-school design, the lines cross approximately -0.5 SD of pretest scores. This indicates that for students with a higher score on the pretest, the POWERSOURCE© intervention is effective. Specifically, in the between-school design approximately 70% of POWERSOURCE© students whose pretest score is higher than -1.0 SD of pretest score mean obtain higher posttest score than students in control groups. Similarly, in the within-school design, POWERSOURCE© students who have higher pretest score have higher posttest score than students in control groups.

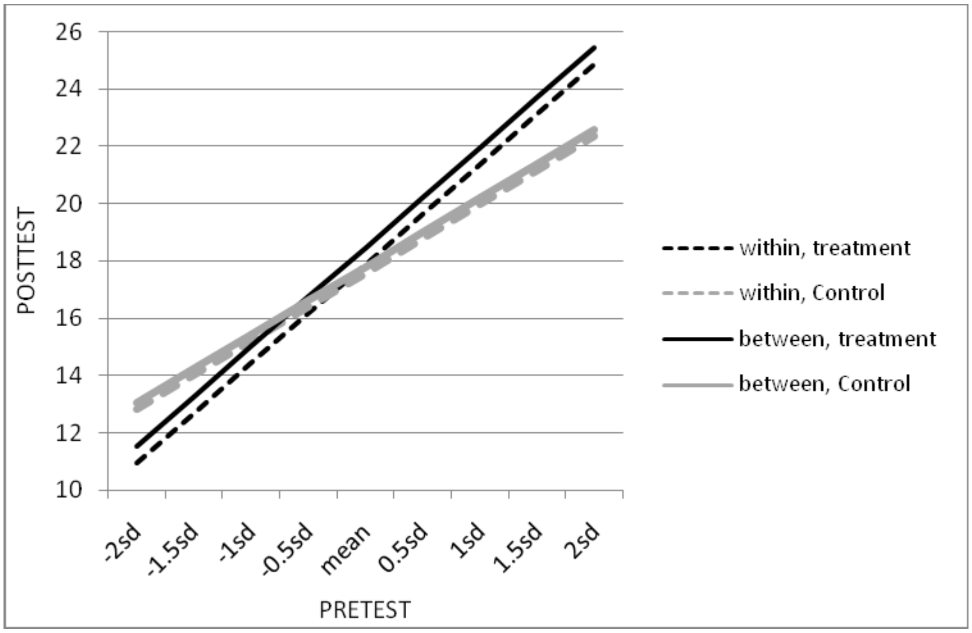


Figure 6. Hierarchical model result. Fitted relationship between pretest and posttest by design and treatment condition.

Sub-domain outcome: PA. The transfer measure contained items relating to all POWERSOURCE© domains. Figure 7 presents the results from items assesssing knowledge of properties of arithmetic (specifically the disitributive property) of which there were five on the transfer measure. On these items we found that the POWERSOURCE© effect is statistically signficant (estimate= 0.57, p-value = 0.002) and the interaction effect between pretest and treatment condition is also significant (estiamte = 0.13, p-value = 0.000).

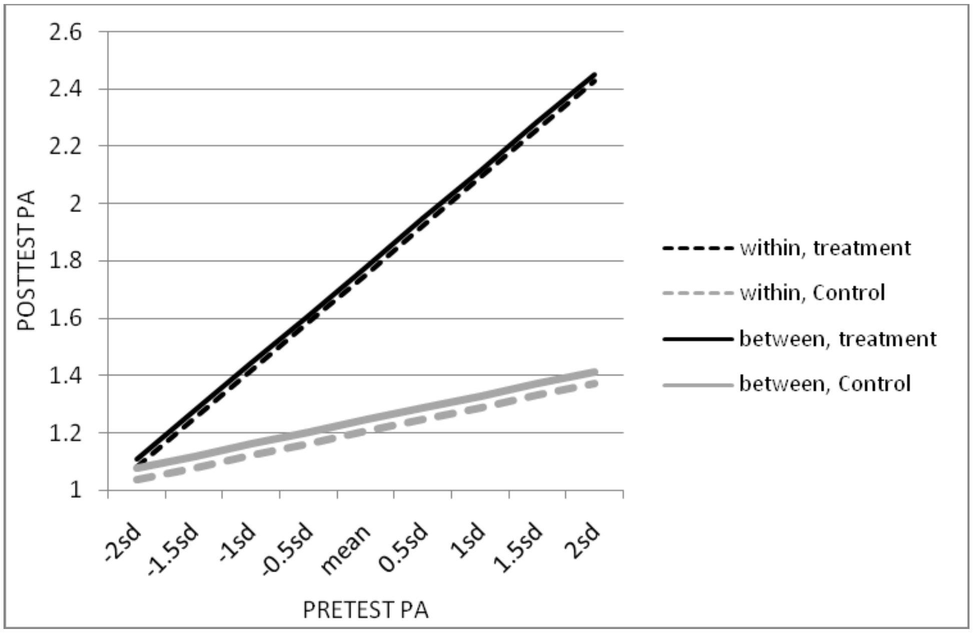


Figure 7. Hierarchical model result. Fitted relationship between pretest and posttest PA subscore by design and treatment condition.

Results from the analyses indicated that:

- A short amount of targeted intervention on key mathematical principles has an impact on student performance on a transfer measure of related content. POWERSOURCE© intervention had more impact on the higher-performing students than the lower-performing students. Those students with higher initial pretest scores tend to benefit more from the treatment when compared to students with lower pretest scores.
- No main effect of design indicating that it did not matter which type of design (within or between schools) we used.
- Transfer measure items related to the properties of arithmetic (specifically the distributive property) were the only ones where we saw a significant effect of POWERSOURCE©. In both designs, students in the POWERSOURCE© group outperformed control group students and the effect was larger as pretest scores increased.
- Item analyses indicated that PA items were more difficult for students that items focused on the other domains Thus, we see a POWERSOURCE© effect on the more difficult items .

Appendix C presents more detailed results for the 2007-2008 implementation year.

Implementation Study 2007-08: professional development and teacher measures. All 7th grade and new 6th grade teachers in the treatment and control groups created a knowledge map and evaluated a sample of student work prior to any professional development. At the end of the school year, after the completion of treatment and control professional development activities, teachers again completed the same knowledge map and evaluation of student work activities.

Of the 113 teachers in our professional development programs, approximately 92 completed some version of a pre- or post-knowledge mapping survey and 86 teachers completed and returned both the pre- and post-evaluation of student work measure. While this year's measures are still being returned as this report is finalized, we provide an analysis of the results of these activities during the 2007 – 2008 school year in the following text.

Teacher Knowledge Maps

To evaluate the maps created by the teachers, each knowledge map was compared to an expert knowledge map created by the researchers conducting professional development. This "expert map" was created by combining the individual maps those researchers created in isolation from one another and from the teachers. Although these individual maps were identical on more than 98% of the relationships and concepts, the researchers met to resolve remaining differences in the map prior to its use as the "expert" standard.

The percentage of agreement between each teacher map and the expert map was analyzed for similarity using five different comparisons. The first comparison measured the exact match between teacher and expert propositions. In this case, a match between teacher and expert requires that two identical concepts be connected using an identical link. For example, Additive Inverse (a concept) is a (link) property of arithmetic (a concept). In the exact match comparison, the "direction" of the link was also specified. For the example just given, "Additive inverse is a property of addition" would be scored as a match, but "A property of addition is a Additive Inverse" would not be scored as a match. In addition to the strictest scoring, we also analyzed the maps without considering the direction of the connection, without the link label, or without either the link label or the direction. In this latter case then, the analysis was just whether two concepts connected by the experts were connected together in any way by the teachers.

To complete the task, the teachers were also asked to connect various problems to the concept map they had created. Specifically, the teachers were asked to link a problem with a concept if that concept was necessary to solve the particular problem. Finally, the teachers were asked to label each problem-concept link with a "2" if knowing the concept alone was sufficient to solve the problem and with a "1 if knowing the concept was necessary, but insufficient by itself to solve the problem. For example, one might be able to solve the problem 12 ÷ 4 by understanding only the concept of division, but understanding only division would be insufficient to find the mean of three numbers.

For the problem-concept part of the mapping task, we analyzed the data in two ways. First, the teacher maps were compared with the experts and rated for exact matches (link and label). Second, the teacher maps were compared with the expert map for any connection (regardless of label) between a problem and concept.

The comparisons were evaluated using a paired t-test to determine the significance of means within treatment and control groups (pre- to post-). A linear regression model was then used to determine the significance of the growth (pre- to post-) between the treatment and control conditions controlling for pre-test score.

Teacher Evaluation of Student Work

All teacher evaluations of student work (both pre- and post-PD) were scored by two raters based on the rubrics developed by Heritage and Vendlinski (2006). When differences in ratings given by these two raters arose, they were discussed and resolved by four other raters to arrive at a "true" score.

The significance of change between the pre- and post-PD evaluations of student work were analyzed within each group using a regression model. In addition, the significance of the changes between the treatment and the control group were analyzed using identical methods. The initial findings from the 2007 – 2008 school year suggests that POWERSOURCE© teachers reconceptualize the mathematics domain associated with understanding algebra and become significantly more "expert-like" compared to the control teachers. On the other hand, the ability of teachers to analyze student work seems to change significantly amongst the POWERSOURCE© teachers in only one area – Rational Number Equivalence. While this could result from the fact that teachers received training on that topic all year, these results could also have occurred by chance since we observed no other changes in teacher ability. An analysis of the data from the end of the 2008 – 2009 school year should allow us to reach more solid conclusions about the effects of our professional development over a long period of time in this area since we can both look for duplication of these results (in the first year 7th grade teachers) and will have teachers that have been exposed to twice the amount of professional development (in the case of the 6th grade teachers). For more detailed findings see Vendlinski, Hemberg, Mundy, Baker, Herman, and Phelan (2009 [Appendix M]).

Teacher Professional Development

POWERSOURCE© teacher Professional Development expanded to include both 6th and 7th grade teachers in the 2008 – 2009 academic year. Many of the sixth-grade teachers are returning from last year, so this part of our Professional Development program focused almost exclusively on how to modify instruction based on the results of student work on our formative assessments. The program of professional development for the 7th grade teachers, on the other hand, blended content knowledge on key, foundational math concepts as they apply in the seventh grade curriculum, on student misconceptions, and on instructional modifications likely to dispel those misconceptions. In three of the seven participating districts, we also provided a program of alternative professional development for teachers not assigned to the POWERSOURCE© treatment. We randomized teachers to groups during their first year of participation, whether in 6th or 7th grade, and these teachers will remain in their respective groups until the end of the study. A brief description of the POWERSOURCE© professional development at each grade is provided immediately below. A description of the Alternative Professional Development programs follows.

The 6th grade professional development meetings are designed with two points of focus; reviewing student response data from the previous POWERSOURCE© unit (completed before the meeting) and utilizing student response data from the previous year of the study to prepare for the upcoming POWERSOURCE© unit. These points of focus were targeted through activities designed around deepening teachers' content knowledge through the analysis of students' responses and response patterns, and through discussions of instructional implications, including teaching strategies. At the beginning of the meeting the teachers were presented with the student response data from the POWERSOURCE© unit that had most recently been completed. This activity has two phases: comparing percentage of correct and incorrect answers overall, and then looking more in depth at response patterns. We would then examine the frequency of correct and incorrect responses, solicit hypotheses from teachers about what these results might mean in terms of student learning, examine response/error patterns to further confirm (or challenge) these hypotheses, then based on these analyses, identify possible instructional responses.

After the student responses had been discussed, the teachers were given a POWERSOURCE© assessment (which corresponded with the upcoming unit) that had been completed using the most frequent incorrect responses from their district (from the previous year) and a worksheet to complete that asked them to identify student errors and potential misconceptions by analyzing the POWERSOURCE© assessment at item level and as a whole assessment. The worksheet asked them to identify various components related to the assessment items, student responses, and potential feedback they would give the student. The worksheet also asked teachers to develop a lesson plan that would effectively teach the unit content and instructional strategies that addressed potential student misconceptions they had identified. The meeting concluded with a review of the actual materials related to the upcoming unit they would be receiving and answering any logistical questions. Primarily attended by teachers who have been in the study for more than 1 year, these professional development meetings focused more on allowing teachers to use their experience and the available student data to think about how to modify instruction to avoid or mitigate misconceptions and student errors. We also wanted to create the possibility for continued peer professional development without facilitation from us by increased teacher involvement and a more collaborative environment.

As was the case for the sixth grade teachers, each 7th grade teacher in the POWERSOURCE© (treatment) group received slightly more than 9 hours of professional development in small clusters (usually between 5 and 20 teachers) by district. These sessions were conducted largely outside of school hours at the district office or at one of the school sites within each district. The initial four hours of PD was almost always done prior to the beginning of the academic year. During this four-hour block, teachers were introduced to the importance of key, foundational topics ("Big Ideas"). The topics are foundational in that much of the content in 6th, 7th, and 8th grade mathematics can be explained and developed from these concepts. In POWERSOURCE©, we focus on three such topics – the multiplicative identity (as applied to Rational Number Equivalence), the meaning of multiplication and other properties of arithmetic (as applied to Distribution), and the meaning of the equal sign (as applied to Solving Equations). The last half of the first session, then focuses on Rational Number (Expression) Equivalence, including proportion with variables. Three 90 minute follow-up sessions with the teachers were conducted in after school settings with the teachers at approximately two-and-a-half month intervals. During the first 45 minutes of each of these follow-on sessions, teachers and researchers discussed student work (from the teachers' students) on the formative assessments associated with a particular foundational concept, possible misconceptions identified by those assessments and possible instructional interventions to correct those misconceptions. The last 45 minutes of each session focused on another single "big idea" (the meaning of multiplication or the meaning of equality) and its application, how that big idea would be developed from its nascent form into abstract concepts in algebra, and how the big idea could be appropriately taught and applied to seventh grade subject matter. To aid teachers with their upcoming instruction on each foundational concept, teachers were given an instructional handbook on that concept during this half of each session. The professional development integrated this instructional handbook (pedagogical content) with the conceptual development of each of the big ideas (content knowledge). Example PowerPoint slides of teacher professional development for both sixth and seventh grade are attached to this report. (Appendix N).

The treatment teachers then returned to their classrooms to develop their actual instructional plan and to provide dedicated instruction to their students on the applicable big idea for two class periods of approximately 40 minutes each. They also administered the *Checks for Understanding* associated with each unit during this time. After the initial presentation of a big idea to their students, teachers were encouraged to continue to use each big idea in other instructional units they developed during the year to teach other concepts.

Teacher Use of Formative Assessment

The use of benchmark technical quality in lieu of formative assessment (Vendlinski & Phelan, 2009 [Appendix O]) and the accompanying paper *Can teacher use of technical quality data about Benchmark Assessments make Benchmarks as effective as Formative Assessments in improving student achievement?* (Vendlinski & Phelan, 2009 [Appendix P]) were presented at the 2009 Annual Meeting of the American Educational Research Association (AERA). Together, these documents outline a comparison of the percentage change of students reaching proficiency from 5th to 6th grade between POWERSOURCE© and Data Director™ schools in one Southern California district. Ostensibly, teachers in the POWERSOURCE© groups were provided materials and training to benefit from formative assessment data. Teachers in the Data Director™ group were only taught how to use results to identify problem areas (student or test related). Our findings suggest that merely training teachers to identify problems may not be as effective as providing teachers training on how to modify their instruction based on formative assessment results.

CCSSO Presentation

In June 2009, we presented similar finding at the National Conference on Student Assessment, sponsored by The Council of Chief State School Officers (CCSSO). This presentation, entitled *Lessons learned: integrating formative, progress monitoring and summative assessment to improve student performance in mathematics* (Schumacker, Vendlinski, & Phelan, 2009 [Appendix Q]) highlighted findings similar to the presentation at AERA, but also conclude that district efforts to improve teacher use of assessment data (POWERSOURCE© and Data Director) are correlated with increased student ability at all levels (not just the levels “proficient” or “advanced”).

POWERSOURCE© Implementation Study 2008-09

The core undertaking of our work during the 2008-09 school year was continuing with an extended, random assignment implementation study of the POWERSOURCE© program. In this year of the study we expanded the intervention from 6th grade only, to 6th and 7th grade in all participating schools. As with prior years, new teachers were randomly assigned to either POWERSOURCE© or control conditions with the ultimate goal of determining program impact on both students and teacher learning outcomes. Teachers continuing in the study for another year maintained their prior year's group status. The 2008-09 study was almost identical to the previous year's work, with a few minor changes:

1. An interim transfer measure was developed for use in Grade 6.
2. We created Grade 7 teacher instructional materials and *Checks for Understanding* Assessments.
3. We modified the professional development sessions (in grade 6) to focus more on interpreting student assessment data and less on teaching the big ideas.
4. We recruited an additional school district to replace a district not continuing with the study.

In the following text, we summarize changes made for the treatment and comparison conditions for the 2008-09 implementation study, including the alternative professional development offered to the control teachers, followed by brief descriptions of the design, measures and analysis plan for the study. Additional details about the plan and its rationale can be found in the supplemental design report submitted to IES in August, 2007. The data collection for these activities are in the final stages.

Development of Grade 7 Materials

During the 2007-2008 year we pilot tested a set of new 7th grade assessment items. Working with expert math teachers, and using data from the 6th grade *Checks for Understanding* we developed a set of 7th grade items reflecting the three conceptual POWERSOURCE© domains culled from our big ideas list—rational number equivalence, principles for solving equations, the distributive property and applications of these big ideas in other critical areas of mathematics. Items were reviewed by a group of five experienced middle-school math teachers.

From the set of 7th Grade items piloted in the 2007-2008 year, we will choose items to include on our *Checks for Understanding* forms and instructional materials for the extension of the POWERSOURCE© study in 7th grade. Items were analyzed using the same procedures outlined in previous reports (Baker, 2008). Several criteria were used to evaluate the items used in the pilot-testing phase. These include: confirmatory factor analyses, reliability analyses, and IRT analyses.

7th Grade Instructional Materials Development

Concurrent to the development of the *Checks for Understanding* items in 7th Grade, we developed instructional materials to be used by teachers. We designed these materials for teachers to use as support when teaching each of the domains addressed in the study. Working with the expert teachers from one of our participating districts, we have developed four Teacher Handbooks—each one closely aligned with the *Checks for Understanding* items in each domain (rational number equivalence, principles for solving equations, the distributive property, review and applications). Knowledge from teaching experience, research on teaching in these areas, and information gathered during the pilot testing year all played a role in developing these instructional materials.

Professional Development 2008-2009

As is suggested in the earlier text, our professional development efforts in the 2008 – 2009 school year doubled over previous years because of the addition of seventh grade teachers to our efforts. In addition, the focus of our efforts with the sixth grade teachers changed significantly. This year's professional development activities in the seventh grade mimicked similar activities with the sixth grade teachers last year. In POWERSOURCE©, the seventh grade teachers were exposed to the notion that certain foundational principles organize other concepts in the math domain and repeatedly occur during the course of instruction on numerous topics. While the foundational principles are the same as they were at sixth grade, such an organization was new to many seventh grade teachers. In addition, we modified our content component so seventh grade teachers could see how these foundational principles applied to the content they taught. We also modified the pedagogical component from last year to account for differences in solution strategies and the different misconceptions evidenced by seventh grade students.

Given that the sixth grade teachers had practiced teaching and assessing the foundational principles the previous school year, we also adapted their professional development to focus more on actually using formative assessment data to formulate instructional strategies to both initially teach and, when necessary, to re-teach the application of big ideas.

As described in the following text, similar adjustments were made to the sixth and seventh grade alternative professional development sessions we designed and implemented.

Alternative Professional Development

In addition to POWERSOURCE© Professional Development, we continued to develop and deliver two alternative types of professional development. These alternative sessions were similar to the alternatives developed for the 2007 – 2008 school year, but were modified to include material relevant to 7th grade teachers. In one district, teachers received instruction in determining the technical quality of district benchmark assessments using a data reporting and analysis tool (Data Director™), and in four other districts, teachers received instruction in student self-efficacy and motivation. Each of those sessions is described in more detail in the following text:

Data director. Teachers using the Data Director™ system to analyze district benchmark tests received approximately 9 hours of professional development during the school year. These sessions were presented in contexts that were nearly identical to the POWERSOURCE© training provided to district teachers, described earlier. The sessions were delivered shortly after the district administered each of three benchmark assessments. While both 6th and 7th grade teachers received similar training, separate sessions were presented to each grade level so that items specific to each benchmark could be discussed. The initial training session was delivered before students took any benchmark tests. This allowed teachers to learn the Data Director™ system as well as an opportunity for these teachers to receive instruction on student performance and technical quality indicators. Among other indicators, the teachers explored measures of central tendency, confidence intervals, inter-item reliability measures, point-biserials, p-values, and a discussion on missing data. The training was conducted using results from the last benchmark of the previous year. In each of the three subsequent 90-minutes sessions, the teachers looked at overall district data, as well as data from their own classrooms to reach conclusions on: 1) how the benchmark performed overall; 2) what items did not perform as expected; and 3) what item(s) students struggled with. It should be noted, that the researchers did not encourage teachers to discuss strategies to (re)teach specific content identified by the teachers as problematic for their students.

Student motivation and self efficacy. We continued with the four themes addressed during the initial year of the program (see Baker, 2008). After an initial extended overview session at the beginning of the year, we met with the teachers for three additional sessions during the year, for approximately 90 minutes. The four theme areas addressed were self-regulated learning, self-efficacy and goals, attributions and affect, and teacher and classroom influences.

The first year's sessions provided general overviews of each of the topic areas, which included exercises to gauge prior beliefs and understanding of how the research in these areas can influence difficulties they may experience in the classroom. The second year entailed a more focused examination of a small number of important research areas within these themes. Those research areas, by theme area, were as follows:

1. Self-Regulated Learning
    a. Behaviors that lead to expertise
    b. Automated vs. controlled information processing
    c. The influence of metacognition on transfer
2. Self-Efficacy and Goals
    a. The effect of modeling on self-efficacy
    b. The effect of progress feedback on goal setting
    c. Mastery vs. performance goal orientation
3. Attributions and Affect
    a. Researched components of anxiety
    b. Differences in mathematics-specific anxiety
    c. Pathways through which emotions affect learning
4. Teacher and Classroom Influences
    a. Classroom scenarios involving the first three focus areas (above)
    b. Teacher feedback on selected approaches to scenarios

The final session entailed having the teachers provide anonymous responses to scenario-based questions to gauge their retention and application of theories discussed during the course of the year. This provided both the facilitator and the teachers with clear indications of how their practices have been shaped through the professional development process.

As in the previous year, our objectives were to introduce teachers to research-based theories and to highlight their connections to everyday classroom practices. In addition to engaging in discussions on the topic areas, we engaged the teachers in brief activities to add practical structure to the application of the theories to daily practice. Attached are some examples of activities and presentations.

In two districts, the alternative condition was "business as usual." While this was intended to be district developed professional development, the researchers did not monitor such professional development activities.

Website Resources

A website has been created to provide participating teachers with resources that assist and enhance their experience while participating in the POWERSOURCE© study. Users of the website will be able to access information and materials that range from logistical information to the research behind the study content. Upon entering the site users are presented with a brief overview of the POWERSOURCE© study and links to download study background and implementation surveys, and a content map of the three Big Ideas. Users are also given the option to view 1 of 4 portals representing the units of the study.

Having direct access to materials and resources on demand provides more flexibility to POWERSOURCE© users and decreases participant's level of dependence on us for materials. This along with a more collaborative professional development setting creates the possibility of a sustainable professional development program within participating districts. The website address is http://www.cresstpowersource.com and members can access the site by entering 203 as an identification number.

Sample and Design

Six districts participated in the random assignment implementation study in 2008-2009. As described earlier we used two designs (within and between school) based on district needs and configuration. Ultimately, three of the districts used a between-school design and three districts used a within-school design. The total number of participants in the study in 2008-2009 are shown in Tables 7 and 8.

Table 7

Sample Distribution by School District ('08-'09 school year) Grade 6

District	*N* of students	*N* of teachers	*N* of schools	Design
AZ-1	590	9	3	BS
CA-1	1225	16	3	WS
CA-2	805	7	2	WS
CA-3	245	7	4	BS
CA-6	1727	33	9	BS
CA-7	170	3	3	WS

Table 8

Sample Distribution by School District ('08-'09 school year) Grade 7

District	N of students	N of teachers	N of schools	Design
AZ-1	355	6	3	BS
CA-1	1310	10	3	WS
CA-2	620	7	2	WS
CA-3	205	4	2	WS
CA-6	1787	23	9	BS
CA-7	590	13	9	WS

Equating Study

For the purpose of equating the scores of various 2008-2009 POWERSOURCE© test forms, the same approach used for the 2007-2008 forms will be applied to the new data.

IRT will be used to equate new test forms. In the context of equating, the new form is often comprised of both new and old operational items where the old operational items, called common items, are previously administered in another form of the test, referred to as the old form.

Some modifications (noted above) were made to the 6th grade transfer measure. Items on the 6th grade *Checks for Understanding* and pretest were the same as the items used in the previous year. Taking those same items into account as common items, the NEAT design will be applied to conduct IPD checking, item parameter linking and test score equating studies. The results will enable us not only to place the scores from 2008-2009 POWERSOURCE© test forms onto the common metric but also to maintain the POWERSOURCE© scales across years.

In 7th grade, the absence of common items across assessment forms requires another external anchor form (EAF) for the 2008-2009 7th grade test forms for the purpose of IRT linking and equating purposes. Based on the NEAT design, FIPC linking, and IRT true-score equating techniques, all test scores from 2008-2009 POWERSOURCE© test forms will be transformed onto the POWERSOURCE© score scale. By including some 6th grade items (about 6 or 7 items) into the new EAF, a vertical test score equating (i.e., across 6th grade and 7th grade test forms) will be attempted.

Transfer Measure

In the 2008-2009 year, the treatment group students in our POWERSOURCE© study received instruction and formative assessments (*Checks for Understanding*) on the four POWERSOURCE© domains. Also included in the study were a control group of students who received their regular instruction.

We hypothesize that students in the POWERSOURCE© group would possess a better understanding of the basic mathematical principles contained within each domain. We also hypothesize that students will be able to apply concepts they have learned, solve complex problems and transfer the principles covered by the POWERSOURCE© domains. For example, having received instruction and formative assessment on rational number equivalence, students should understand the multiplicative identity principle and be able to use it to: a) demonstrate that a set of rational numbers are equivalent, b) find equivalent fractions, c) find missing numbers in proportions and d) solve proportional reasoning problems? In order to answer these questions we used a transfer measure (posttest) to compare the POWERSOURCE© and Control groups on novel items related to our four POWERSOURCE© domains.

Grade 6 Transfer Measure

The Grade 6 Transfer measure was first used in 2007-2008. This transfer measure was developed using items from several sources including TIMMS, NAEP, the QCA Key Stage 3 exam, PISA and Benchmark tests used in one of our pilot districts (see Appendix R for sources of all items). An initial set of 44 items were selected from the various sources. Items were selected based on their relevance to the POWERSOURCE© domains and their appropriateness for a transfer task (related to POWERSOURCE© content, but not exact replicas of item types used in the *Checks for Understanding*). A final set of items (29) were selected from the initial 44 items. Of these items 19 were multiple choice, 9 short answer and 1 explanation task. Items were selected based on their representation in the CA state standards and relevance to POWERSOURCE© items (see Appendix S for the alignment of items to CA standards and the NCTM Focal points). Some of the initially selected items were deemed more appropriate for 7th grade and were used for the 7th grade transfer measure.

Grade 6 Transfer Measure Revision

Based on the item analyses outlined in the earlier text, we decided to modify the Grade 6 transfer measure to decrease the number of multiple choice items, and increase the number of extended response items. Because the amount of information we get from an extended response item is so much greater than for a multiple choice item, the more extended response items on the test, the fewer multiple choice items are required. We looked at the 95% confidence interval of the transfer measure items in order to determine which items were of equal difficulty. We found 5 pairs of items each with identical or overlapping confidence intervals. We eliminated one of each of these pairs. We next modified some of the existing item formats from multiple choice to extended response. For example, in the original transfer measure we had an item asking students to write a different fraction equivalent to 3/5. In the revised measure, we asked students to find the fraction and then, in a second part to the question, to explain why the two fractions were equivalent. Items were also reordered to reflect the difficulty patterns seen in the item analyses—from easiest to the most difficult (see Appendix T for an updated list of Grade 6 Transfer Measure items).

Grade 6 Interim Transfer Measure

In an effort to gather more student outcome data, we designed an interim transfer measure to be given to students after completion of the first two POWERSOURCE© domains. These domains were PA and RNE. We created a 20 item test form with 20% of the items requiring students to explain a concept in their answer. We selected two items per domain from the pretest (of medium difficulty) and changed the numbers in the items. The remaining items were taken from the transfer measure and again were modified to include different numbers, and/or situations. Items selected for the interim transfer measure had a range of difficulty from b=1.101, p-value = .17, to b= -1.441, p-value = 0.88 (see Appendix U for Grade 6 Interim Transfer Measure).

Grade 7 Transfer Measure

The Grade 7 Transfer measure was developed using similar procedures as the Grade 6 transfer measure. Items were selected from TIMMS, NAEP, the QCA Key Stage 3 exam, PISA and Benchmark tests used in one of our pilot districts (see Appendix V for sources of all items). Items were selected based on their relevance to the POWERSOURCE© domains and their appropriateness for a transfer task (related to POWERSOURCE© content, but not exact replicas of item types used in the *Checks for Understanding*). An initial set of 51 items were selected and narrowed down to a final pool of 26 items. Of these items 17 were multiple choice and the rest were either short answer or explanation tasks, or a combination of both types. Items were selected based on their representation in the CA state standards and relevance to POWERSOURCE© items (see Appendix W for the grade 7 Transfer Measure).

Observation and Interview Study

As part of the 2008-09 POWERSOURCE© implementation study we conducted a study of classroom observations and teacher interviews. This followed a pilot study in 2007-

2008 of the interview and observation measures. These observations/interviews had several inter-related purposes: First, they provided first-hand data, to supplement the self-report surveys about how teachers were using POWERSOURCE© materials in the classroom, including assessments, instructional activities, and learning supports. Second, they provided a more open-ended opportunity for teachers to provide feedback about their POWERSOURCE© implementation and professional development experiences. Finally, it allowed us to pilot instruments and methodology for scaled up qualitative data collection in the remaining years of the study.

Methodology

Ten 6th grade teachers were observed and interviewed about their use and implementation of POWERSOURCE© during spring, 2008 (one additional teacher participated in interviews but, due to scheduling issues, was not observed in the classroom). The teachers were from four of the districts participating in the larger POWERSOURCE© study.

Teachers were observed during the final POWERSOURCE© unit, RA (Review and Applications) using a semi-structured observation protocol (see Appendix X). The RA unit was selected because it incorporates all of the big ideas covered throughout the POWERSOURCE©, and also because teachers have experienced all aspects of the POWERSOURCE© professional development and had access to all POWERSOURCE© resources by the time of completion of this unit. Each teacher was observed for one classroom period by two observers. All observers participated in a 4-hour training for protocol use, and were trained to concordance on the protocol's rating scales.

One-on-one interviews were conducted with each teacher after the observation. Interviews used a semi-structured protocol and were audio-taped and transcribed. Data was coded and analyzed using qualitative data software (AtlasTi). Coders were provided a priori codes and trained to concordance before the coding process began. Each interview was coded independently by two coders.

Results

In the following text we summarize some of the key trends found across this qualitative data set. Given the small sample and pilot nature of this first qualitative data collection endeavor, caution should be taken in generalizing results to the sample as a whole. However, the findings provided some useful information and insights as the implementation continues to move into other schools or districts. For the purposes of presentation, the discussion is divided into teacher implementation/use of the POWERSOURCE© materials and teacher feedback regarding POWERSOURCE© participation.

Teacher feedback. All of the teachers provided positive feedback about their participation in POWERSOURCE© and use of the POWERSOURCE© materials. Teachers commented that the materials were easy to use, quality, and helpful augmentation to their instruction. Teachers highlighted the role of the "big ideas" of algebra, and how they could be translated into practice, as having a positive impact on both their conceptual math understanding and instructional practice. Teachers specifically compared the POWERSOURCE© materials to their regular district curriculum, indicting the instructional methods were different than what was typical of their district curriculum, and that the majority of them also preferred the POWERSOURCE© approach.

Even with these differences, the teachers felt that there was good alignment (four teachers used the words "perfect" or "beautifully") between the materials and district standards. When concerns were raised about integration of POWERSOURCE© with district curriculum they focused on pacing (i.e., fitting POWERSOURCE© time wise within the district curriculum) rather then on content/standards. The teachers also noted that preparation for a POWERSOURCE© lesson took the same or slightly more time than preparation for a lesson from their regular math curriculum, so did not present an undue burden on their instructional and planning time.

In terms of the POWERSOURCE© professional development, all of the teachers were able to identify and describe useful concepts and strategies gained from the POWERSOURCE© professional development that they used in their classroom. Although the concepts/content identified crossed all of the units (RNE, PA, SE, RA), singled out in particular were the concept of a number over itself being equal to 1 (i.e., "The Big One"), and the area/array model for repeated addition. The teachers reported thinking that the amount/level of professional development provided for POWERSOURCE© adequately prepared them to implement the program – that is, they were able to grasp the concepts and strategies needed to implement the program even with the modest number of hours for POWERSOURCE© direct professional development delivery.

A subset of teachers suggested that the professional development could be improved by reducing the time spent of reviewing prior student results, and more time spent on covering the big ideas and instructional strategies that provide the foundation for POWERSOURCE©. Although their other responses indicate that they value the formative review of student data, this specific suggestion appeared to spring from the teachers' desire to increase their conceptual understanding of the foundations of algebra, which had perhaps not been a specific focus of prior in-service or pre-service professional development they had previously participated in.

Teacher implementation/use of POWERSOURCE© materials. Based on both interview and observation, teachers appeared to adhere fairly closely to the teacher handbooks lessons/provided resources in their implementation of the POWERSOURCE© lessons. It is emphasized throughout the professional development that the handbook is a guide or resource and that teachers do not need to follow it explicitly. The observed teachers, however, appeared to prefer to not vary from the provided script, which may be due to the fact that POWERSOURCE© is their first exposure to many of the core algebraic concepts presented. In the interviews teachers reported integrating some of the POWERSOURCE© instructional concepts or ideas (such as the aforementioned “Big One” and array model) into regular math lessons.

Both the interviews and observations provided evidence that the teachers were using the *Checks for Understanding* as designed. Teachers reported reviewing the information from the *Checks* to help make instructional decisions and guide their follow up lessons, such as identifying concepts that need to be reinforced with their students, deciding what types of additional practice problems to provide, and how to group students for POWERSOURCE© lesson 2 (consistent with this report, a variety of groups strategies and patterns were observed in the classrooms). At the same time, consistent with the formative underpinnings of the POWERSOURCE© model, all but two of the teachers were observed repeatedly both questioning students for understanding and using guided problem solving to help students find the correct problem solutions.

There were also some trends in terms of what teachers reported that they found when they reviewed the *Checks*. For example, the most common areas of “weakness” or instructional need identified by teachers based on their review of the *Checks* was concepts related to the properties of arithmetic, predominantly the distributive property. On the other side, concepts related to solving equations and using ratios were the areas of student strength most identified by the teachers. Another area of difficultly worth noting relates to student familiarity with specific item formats. That is, some teachers reported that students had some initial confusion with the item formats used in the *Checks*, such as explanation tasks, showing their work, or apply concepts and responses to different boxes or shapes. In other words, these types of items were atypical of the multiple choice/fill in items that students are regularly asked to complete, and there was a learning curve in terms of getting students familiar with answering these different types of questions. These reported difficulties also suggest that student may begin POWERSOURCE© with a limited ability to transfer what they know to new settings and contexts, and that this is an additional instructional support that teachers may need to consider in their implementation process.

Next Steps

The qualitative data collection process was expanded for the 2008-09 school year. We targeted to double the number of teacher participants to 20, which would represent 26% of the total teachers who implemented POWERSOURCE© during the 2008-09 school year. As of this writing, 16 teachers from 5 of the POWERSOURCE© districts have participated in the 2008-09 teacher interviews and observations. The majority of these subjects (10) participated in POWERSOURCE© for at least one year prior. The scope of the observations themselves have also been increased, with researchers observing two days of lessons for two units (SE and RA) for each of the participant teachers.

For the upcoming school year (2009-10) we plan to target a sample of 20-25 teachers for interview and observation. Based on on-going logistical difficulties with coordinating observations towards the end of the school year the two units we plan to observe are PA and SE, with PA replacing RA (which is typically implemented towards the end of the school year) in the observation plan. As described in the earlier text, PA and SE represent the two areas where the 2007-08 interview sample felt their students were the strongest (SE) and the weakest (PA). The observation and interview protocols will be revised to address this change.

Analysis Plan

Plans for analyzing data from the 2008-2009 year include:

- POWERSOURCE© has implemented in grades 6 and 7 during the 2008-2009 school year. Since the study has implemented a similar design and instrumentation described earlier, we will basically utilize a similar statistical models and analyses plan as employed in 2007-2008 study. Note, however, that the analyses will be conducted separately by grade level, i.e., one set of analyses for data in the 6th grade level and the other set of analyses for data in the 7th grade level.
- One of the key distinctions for the 2008-2009 data analyses is that we will explore some possibilities of examining student growth trajectory during a year with three time-series measures: pretest, interim transfer measure, post transfer measure. We will address the following interesting questions: what does the growth trajectory look like?; how much variability in student growth trajectory is observed?; Does the rate of growth differ between the control group and the POWERSOURCE© group?
- Given this is the second year of the POWERSOURCE© large-scale implementation, we are keenly interested in differential/cumulative effects of the

POWERSOURCE© experience both in students and teachers. For example, we expect there to be a significant impact of number of years a teacher has been involved in POWERSOURCE©. We hope to see that teachers will become more proficient in their subject matter knowledge, more skilled in their formative use of assessment, and better equipped to focus their instruction on key ideas. And, as a result, teachers will be more effective in helping students to improve their understanding of key algebra principles.

- The focus for project implementation during the 2009-10 school year will be continuing the experimental (random assignment) study of POWERSOURCE© impact begun in 2008-09. Specifically, in addition to continuing the study at the 6th and 7th grade levels, we plan to add the 8th grade classrooms in the participating districts to the study.

8th Grade Materials Development and Pilot Testing

In addition to the implementation study at grade 6 and 7, during the 2008-09 school year we developed and pilot tested materials for 8th grade POWERSOURCE© modules, towards the goal of adding 8th grade teachers to the experimental study in 2009-10. This work included the development and testing of 8th grade *Checks for Understanding* as well as the development of instructional and professional development support materials.

Development of 8th Grade Items

During the 2008-2009 year we pilot tested a set of new 8th grade assessment items. Working with expert math teachers, and using data from the 6th, and 7th grade *Checks for Understanding* we developed a set of 8th grade items reflecting the three conceptual POWERSOURCE© domains culled from our big ideas list—RNE, principles for solving equations, the distributive property and applications of these big ideas in other critical areas of mathematics. Items were reviewed by math experts as well as a group of five experienced middle-school math teachers.

The three conceptual domains chosen for inclusion in the POWERSOURCE© study were selected because they are: a) heavily represented in state standards and state and district test blueprints; b) historically difficult for students to master; and c) important prerequisites for learning and mastering algebra. Given that the conceptual domains for POWERSOURCE© remain constant across the grade levels, we paid close attention to how the concepts develop across grades. Thus, in developing the 8th grade assessment items, we looked at the learning trajectory for each domain. In 6th grade students were learning how to solve one-step linear equations, 7th grade they begin to solve more complex, two-step equations, and in 8th grade students are simplifying expressions, solving linear equations and solving multi-step problems (including word problems) . The big ideas underlying the concepts remain the same, but the skills students are mastering are different and build on what was learned previously. Our POWERSOURCE© materials and *Checks for Understanding* assessments reflect this trajectory and make the necessary connections between the 6th and 7th grade content and the big ideas (see Appendix Y for an example of the CA state standards relating to solving equations).

Pilot Testing of 8th Grade Items

Around 75 8th grade items have been developed and 40 items pilot tested 11 teachers in 3 schools. Using the same assessment model as the 6th and 7th Grade items, we have developed different types of assessment: basic computation tasks, partially worked problems, explanation tasks, word problems and problems involving graphics. Items were grouped together (within domains) to create the *Checks for Understanding* assessment forms. We used an overlapping design to allow us to compile item data and conduct IRT analyses on all items. The items we have pilot tested to date were compiled into 14 forms.

Pilot testing process. For pilot testing, the tasks described were assembled into forms that students should be able to complete in about 15 minutes. This time frame was imposed by the districts we were working with. Any assessment longer than that, they felt, would be seen by teachers as a test, and would evoke complaints about too much district testing. However as it has turned out, this time frame actually has a number of advantages in focusing teachers and students' attention on students' understanding of a single concept and encouraging deep assessment without being too intrusive into or engendering teacher hostility about intrusion into instructional time.

Each teacher participating in pilot tests received at least two different test forms, each focusing on the same big idea, with each form containing between 3-5 tasks. The forms were randomly assigned to students within classrooms, and each teacher administered the assessments to all of their 8th grade students. In all cases the first 2-3 items on the test forms were basic computation items. The subsequent items were open-ended explanation tasks, partially worked problems, word problems, or problems with a graphic prompt. Forms containing explanation tasks did not contain any other tasks besides the basic computational items.

All pilot data from the closed-ended responses were entered and by a group of undergraduate and graduate student workers and other CRESST staff. Three-point scoring rubrics were developed for the open-ended items.

Selecting items for inclusion in the 8th grade *Checks for Understanding*

From the set of 8th Grade items piloted in the 2007-2008 year, we will choose items to include on our *Checks for Understanding* forms and instructional materials for the extension of the POWERSOURCE© study in 8th grade (to be conducted in the 2009-2010 school year). Data from the pilot test are currently being analyzed and we will use the same procedures for analyzing data and selecting items as we used for the 6th and 7th Grade *Checks for Understanding*. That is, as indicated earlier, we will employ several criteria to evaluate the items used in the pilot-testing phase. These include: confirmatory factor analyses, reliability analyses, and IRT analyses. Specifically, our typical analysis scheme for each extended set of *Checks for Understanding* from each pilot test form was to first calculate reliability coefficients (Cronbach's alpha) for items representing each domain. Second, as another check of item quality, we conducted a principal component analysis and a confirmatory factor analysis for each test form to check whether the items exhibited the factor structure we expected; e. g., whether the computation items loaded on the same factor, etc. Third, IRT analyses based on Rasch models were conducted in order to obtain item parameters (difficulties) and item characteristic and information curves so that we could use them to select items for future testing. The model-data fit was investigated using two model fit indices. One is the G2 index which is the Chi-square ($\chi 2$) statistic and provided in PARSCALE phase 2 outputs, and the other is the MNSQ statistics.

Eighth grade instructional materials and professional development design. Concurrent to the development of the *Checks for Understanding* items in 8th Grade, we are developing instructional materials and professional development supports. To date these materials are in draft form and have been developed with input and advice from five expert middle-school math teachers. As with the 6th, and 7th grade materials, knowledge from teaching experience, research on teaching in these areas, and information gathered during the pilot testing year all play a role in developing these instructional materials.

Supplementary Research Activities

Following is a brief update of a supplementary strand of work undertaken as part of the Center activities during the 2008-09 school year. This work includes an investigation of investigation of district contexts for assessment.

Use of Interim Assessment Data/District Contexts

This research activity takes a broader contextual approach to interim assessment use, examining the ways in which middle school mathematics teachers use the data provided by POWERSOURCE© and other types of interim assessments, and how the features of the assessments are related to data use. The project is being conducted simultaneously in three sites—Central Colorado (coordinated by Lorrie Shepard, CU Bolder), Southern California (coordinated by Brian Stecher, RAND), and Northern California (coordinated by Hilda Borko, Stanford). We selected districts that had invested in teacher professional development around formative assessment or had installed formal interim assessment systems.

During the past project year we have interviewed 26 administrators in ten school districts in three locations: four districts in Central Colorado, four in Southern California and two in Northern California. From those districts, we selected 18 schools with middle grades (6, 7 or 8), and we interviewed the principal or assistant principal in each school. In addition, we interviewed 42 middle grades mathematics teachers in those 18 schools. Each teacher was interviewed twice. The first interview was usually conducted by telephone; the interview allowed us to obtain preliminary information about assessment practices and to instruct teachers how to collect artifacts in advance of the second, in-person interview. The second interview went into greater depth about the nature of assessments occurring in the classroom and the teacher's use of information obtained from the assessments. This interview was structured around assessment artifacts collected by the teacher, and we were given de-identified copies of the artifacts to use in our analyses. We conducted approximately half of the interviews during the spring of 2008, and the rest in the fall and winter of 2008/2009. All interviews were recorded, and all the audio recordings have been transcribed and imported into Nvivo 8.0 for analysis. In addition, all the artifacts have been scanned. The files are accessible to research team members on a Sharepoint site maintained by RAND.

The data analysis is being conducted in several phases. Initially we read through the teacher transcripts in an unstructured way, looking for themes and patterns of responses and thinking about ways to approach the coding process. Next, individual researchers formulated thematic codes based on their readings. The research team held extended discussion about the codes, creating definitions with exemplary quotes from the interviews. The team also formulated a structure for grouping codes under a set of broader categories, including the nature of the assessments, the kinds of information that were provided, how the assessment data were used to modify classroom practices, teachers' attitudes toward assessments, and other general topics. Each of these categories admitted five to ten detailed subcodes. We assembled a "code book" and tested it by having all researchers code the same interview using nVivo. We compared codes, discussed differences, made some modifications to the code book, and added some coding conventions. We conducted additional rounds of coding until we were satisfied that team members shared a common understanding of the code book and coding conventions. The all interviews were assigned to pairs of researchers, who coded the files in nVivo and reconciled any discrepancies. We completed the first round of coding in April 2009. We are using these codes to generate targeted "reports" around important themes, which will be the basis for subsequent reading and analysis. This process was just begun in May 2009, along with a separate reading of the transcripts from the districts and school interviews. Analysis will continue through the summer and fall, and we will begin writing reports in the winter of 2009.

Leadership

A core planned set of supplemental activities is our leadership strand of work. Our leadership activities intend to support states and districts in their desire to develop coherent instructional programs to engage in standards-based reform. The work focuses in two areas. First, it will focus on the collaborative development of methodology and annotated examples that practitioners and contractors can use to align instruction and assessment developmentally with key priorities for student capability in mathematics as well as with standards. The methodology seeks deeper understanding and communication of the learning demands inherent standards and the developmental progressions that are essential to accomplishing key standards. The methodology lays out a systematic framework describing these learning demands and progression, rather than simply working backward from one existing test. Products from the proposed effort will include software with embedded tutorials for conducting alignment analyses, paper and poster illustrations, and the results of workshops and webinars held with experts in math, math education, test developers, and other researchers as well as with the practitioner and policy communities.

Formative Assessment Group

Recently, several CRESST researchers have formed a working group to define assessment quality as it applies in its broadest sense to formative assessment. While there exists a growing body of empirical work on the benefits of formative assessment to student learning (e.g., Black, Harrison, Lee, Marshall, & Wiliam, 2004; Black &Wiliam, 1998; Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002), this literature has mainly addressed the process of formative assessment. The formative assessment process is characterized as continuous, carried out during the course of teaching and learning to provide feedback to teachers and students to improve teaching and learning. Discussions of assessment quality are less prominent in the formative assessment literature. The goal of our CRESST working group is establish a framework for considering formative assessment quality.

Prior work (Phelan, et al., 2009) has shown us that we can establish technical quality of formative assessments and data suggest that relatively brief formative assessments focused on key conceptual domains can provide reliable and useful information on students' levels of understanding and possible misunderstandings in the domain. These results, however, are just part of the evidence needed to validate the tasks as formative assessments. Other evidence includes information on the sensitivity of the tasks to instruction (so that they are not just measuring, for example, general intelligence or mathematics achievement) and the utility of the tasks in a formative assessment system, which means that teachers are able to use the assessments to make more informed and effective instructional decisions.

Formative assessment can include questioning, discussions, tasks, representations, and explanations. Whatever the assessment strategy, formative assessment is not "formative" unless action is taken on the basis of the evidence the assessment provides. The action is intended to lead to further learning and thus to have positive consequences (e.g., Moss, 2003; Stobart, 2006). However, positive consequences hinge directly on teachers abilities to interpret the evidence and to know what action to take as a result. Effectively interpreting and using evidence is dependent on teacher knowledge: domain knowledge and pedagogical content knowledge. As a step toward developing an assessment quality framework, our working group is currently engaged in analyzing the range of teacher knowledge needed for different types of formative assessment. A structure for our analysis is shown in Table 9.

Table 9

Structure for Analyzing Teacher Knowledge

Assessment cycle	Cognitive demand	Formative assessment	Type of evidence	Teacher knowledge	Teacher action
Length of the assessment cycle – e.g., 5 minutes, 1 lesson, 1week	Cognitive demand of the assessment task	Example of a formative assessment linked to cycle and cognitive demand	Evidence provided from the formative assessment	Knowledge needed to interpret the evidence i.e. what does this tell me about current learning status?	Desirable action to move learning forward

Although we are in the early stages of this analysis, we anticipate it will yield insights into some key considerations of assessment quality, which will inform the next step of our work toward establishing a framework for assessment quality related to formative assessment.

Plans for 2009-2010

Currently, we are beginning data analysis of data collected during the 2008-09 school year, including the student transfer measures, pretest measures and the multiple teacher outcomes described in this paper. Additionally, we will analyze the *Checks* completed by the POWERSOURCE© group teachers, both in terms of statistical quality of the items and to track student scores across the school year. We will also analyze state test data outcomes as they are made available by the districts including, when available, subscale scores of state mathematics items.

The focus for project implementation during the 2008-09 school year will be continuing the experimental (random assignment) study of POWERSOURCE© impact begun in 2008-09. Specifically, in addition to continuing the study at the 6th and 7th grade levels, we plan to add the 8th grade teachers in the participating districts to the study (note that, depending on district configuration, there may be some overlap in sample in cases where the same teachers teach both multiple grades of math). The study will utilize a similar design and instrumentation to that described in the earlier text regarding the 2008-09 study, with student and teacher outcome instruments adapted to reflect 8th grade content as applicable.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19, (6),* 716-723.

Atkinson, R.C., & Shiffrin, R.M. (1968) Human memory: A proposed system and its control processes. In K.W. Spence and J.T. Spence (Eds.), *The psychology of learning and motivation, vol. 8.* London: Academic Press Bock, R., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275-285.

Baker, E.L. (2006, June). National Center for Research on Evaluation, Standards and Student Testing. The development and impact of Powersource, June 1, 2006. PR/Award #: R305A050004.

Baker, E.L. (2007). The Development and impact of POWERSOURCE© (Progress Report to the Institute for Education Sciences Year 2). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Baker, E.L. (2008, June). National Center for Research on Evaluation, Standards and Student Testing. The development and impact of Powersource, June 1, 2008. PR/Award #: R305A050004.

Ball, D.L., & Bass, H. (2001). What mathematical knowledge is entailed in teaching children to reason mathematically? In National Research Council, *Knowing and learning mathematics for teaching: Proceedings of a workshop* (pp. 26-34). Washington, DC: National Academy Press. Retrieved from http://books.nap.edu/catalog/10050.html

Ball, D.L., Lubienski, S., & Mewborn, D. (2001). Research on teaching mathematics: The unsolved problem of teachers' mathematical knowledge. In V. Richardson (Ed.), *Handbook of research on teaching (4th ed.).* New York: Macmillan.

Black, P., Harrison C., Lee, C., Marshall, B., and Wiliam, D. (2004). Working Inside the Black Box: Assessment for Learning in the Classroom. *Phi Delta Kappan,86,* 9-21.

Black, P.J., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5(1),* 7-74.

Black, P.J., & Wiliam, D. (1998b*). Inside the black box: Raising standards through classroom assessment.* London: School of Education, King's College. (See also article with the same title, 1998, in Phi Delta Kappan, 80(2), 139-148.)

Bloom, B.S. (1968). Learning for mastery. *Evaluation Comment, 1(2)*, 1-12.

Bock, R.D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261-280.

Brown, A., Bransford, J., & Cocking, R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school.* Expanded edition. Washington, DC: National Academy Press.

Brown, R.S., & Niemi, D.N. (2007). *Investigating Alignment of High School and Community College Assessments in California.* San Jose, CA: National Center for Public Policy in Higher Education.

Carpenter, T., & Franke, M. (2001). Developing algebraic reasoning in the elementary school. In H. Chick, K. Stacey, J. Vincent, & J. Vincent (Eds.), Proceedings of the 12th ICMI Study Conference (Vol. 1, pp. 155-162). Melbourne, Australia: The University of Melbourne.

Catrambone, R., & Holyoak, K.J. (1989). Overcoming contextual limitations on problem solving transfer. Journal of Experimental Psychology: Learning, Memory and Cognition, 15(6), 1147-56.

Chi, M.T.H., & Bassok, M. (1989). Learning from examples via self-explanations. In L.B. Resnick (Ed.), *Knowledge, learning, and instruction: Essays in honor of Robert Glaser.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Chi, M. T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5, 121-152.*

Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement*, *27*, 440-458.

Ericsson, K.A. (2003). The search for general abilities and basic capacities: Theoretical implications from the modifiability and complexity of mechanisms mediating expert performance. In R. J. Sternberg & E. L. Grigorenko (Eds.), *Perspectives on the psychology of abilities, competencies, and expertise* (pp. 93-125). Cambridge: Cambridge University Press.

Ericsson, A.K., & Simon, H.A. (1984). Protocol analysis. Verbal reports as data. Cambridge, MA: MIT Press.

Fuchs, L.S., Fuchs, D., Finelli, R., Courey, S.J., & Hamlett, C.L. (2004, Summer). Expanding Schema-Based Transfer Instruction to Help Third Graders Solve Real-Life Mathematical Problems. *American Educational Research Journal, v41 n2* p419-445.

Haverty, L. (1999). *The importance of basic number knowledge to advanced mathematical problem solving.* Doctoral Dissertation, Carnegie Mellon University, Pittsburgh, PA.

Heritage, M., & Vendlinski, T. (2006). Measuring Teachers' Mathematical Knowledge (Technical Report No. 696). Los Angeles: UCLA / CRESST.

Heritage, M., & Yeagley, R., (2005). Data use and school improvement: Challenges and prospects. In Herman, Joan L. and Haertel, Edward H. (Eds.). *Uses and Misuses of Data for Educational Accountability and Improvement.* The 104th Yearbook of the National Society for the Study of Education. Part 2. Malden, Massachusetts and Oxford, England. Blackwell Publishing.

Herman, J. L., & Baker, E. L. (2006). Making benchmark testing work for accountability and improvement: Quality matters. *Educational Leadership, 63(3),* 48-55.

Herman, J. L., & Gribbons, B. (2001). Lessons learned in using data to support school inquiry and continuous improvement: Final report to the Stuart Foundation (CSE Tech. Rep. No. 535). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J., Osmundson, E., Ayala, C., Schneider, S., & Timms, M. (2006). *The Nature and Impact of Teachers' Formative Assessment Practices.* CSE Technical Report 703. National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Paper prepared for the Annual Meeting of the American Educational Research Association (Montreal, Canada, Apr 2005).

Hiebert, J., & Carpenter, T. P. (1992). Learning and teaching with understanding. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning (pp. 65-97).* New York: Macmillan.

Judd, C. H. (1908). The relation of special training to general intelligence. *Educational Review, 36,* 28-42.

Judd, C. H. (1936).*Education as the Cultivation of Higher Mental Processes,* Macmillan, New York.

Kilpatrick, J. (1992). A history of research in mathematics education. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning (pp. 3-38).* New York: Macmillan.

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it Up: Helping Children Learn Mathematics.* Report of the Mathematics Learning Study Committee. National Research Council, National Academy Press: Washington, D.C.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254-284.

Kolen, M.J., & Brennan, R.L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices.* New York: Springer.

Lord, F.M. (1982). Item response theory and equating—A technical summary. In P. W. Holland & D. B. Rubin (Eds.), *Testing Equating* (pp. 141–161). New York: Academic.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Mayer, R. E. (2003). *Learning and instruction.* Upper Saddle River, NJ: Merrill Prentice Hall.

Moreno, R., & Mayer, R. E. (2005, February 5). Role of guidance, reflection, and interactivity in an agent-based multimedia game. *Journal of Educational Psychology, v97 n1* p117-128.

Moss, P.A. (2003). Reconceptualizing validity for classroom assessment.

Muraki, E., & Bock, E.D. (1997). Parscale IRT item analysis and test scoring for rating scale data. Chicago, III: Scientific Software International

Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA: Harvard University Press. No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and monitoring activities. *Cognition and Instruction, 1,* 117-175.

Pawley, D., Ayres, P., Cooper, M., & Sweller, J. (2005). Translating words into equations: A cognitive load theory approach. *Educational Psychology, 25,* 75-97.

Phelan, J., Kang, T., Niemi, D. N., Vendlinski, T., & Choi, K. (2009). Some Aspects of the Technical Quality of Formative Assessments in Middle School Mathematics, CRESST Report 750.

Plake, B. S., & Impara, J. C. (1997). Teacher assessment literacy: What do teachers know about assessment? In G. Phye (Ed.), *Handbook of classroom assessment (pp. 53–68).* San Diego, CA: Academic Press.

Pressley, M., & Brainerd, C. J. (Eds.). (1985). Cognitive learning and memory in children; Progress in cognitive development research, New York: Springer-Verlag.

Ready, T., Edley, Jr., C., & Snow, C. E. (Eds.). (2002). *Achieving High Educational Standards for All: Conference Summary.* Washington, DC: National Academy Press.

Richardson-Klavehn, A., & Bjork, R.A. (2002). *Memory: Long term. Encyclopedia of cognitive science. Vol. 2* (pp. 1096-1105). London: Nature Publishing Group.

Ruiz-Primo, M.A., Shavelson, R.J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching, 39(5*)**,** 369-393.

Schmidt, W. H., McKnight, C.C., & Raizen, S. A. (1997). *A splintered vision: An investigation of U.S. science and mathematics education.* Boston: Kluwer Academic Publishers.

Schumaker, V., Vendlinski, T.P., & Phelan, J. (2009). Lessons learned: Integrating formative, progress monitoring and summative assessment to improve student performance in mathematics. Council of Chief State School Officer's 39th National Conference on Student Assessment. Los Angeles, CA.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461-464.

Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *Handbook of research on teaching, (4th ed., pp. 1066–1101).* Washington, DC: American Educational Research Association.

Stiggins, R. (2005). From Formative Assessment to Assessment FOR Learning: A Path to Success in Standards-Based Schools. *Phi Delta Kappan, Vol. 87, No. 04,* December 2005, pp. 324-328.

Stobart, G. (2006). Influencing classroom assessment. *Assessment in Education: Principals, Policy & Practice, Vol. 12, No.* 3, 235-238.

VanLehn, K. (1996). Cognitive skill acquisition. In J. Spence, J. Darly & D. J. Foss (Eds.), *Annual Review of Psychology, Vol. 42,* pp. 513-539). Palo Alto, CA: Annual Reviews. 4 5.

Vendlinski, T. P., Hemberg, B. C., Mundy, C., Baker, E. L., Herman, J. L. Phelan, J., et. al. (2009). *Designing professional development around key principles and formative assessments to improve teachers' knowledge to teach mathematics.* Meeting of the Society for Research on Educational Effectiveness. Crystal City, VA.

Vendlinski, T.P., & Phelan, J. (2009). The use of benchmark technical quality in lieu of formative assessment. Annual Meeting of the American Educational Research Association, San Diego, CA.

Wolf, D., Bixby, J., Glenn, J. III, & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of Research in Education, Vol. 17,* pp. 31-74). Washington, DC: American Educational Research Association.