Understanding Bias Due to Measures Inherent to Treatments in Systematic Reviews in Education

Robert E. Slavin

Nancy A. Madden

Johns Hopkins University

-and-

University of York

Paper presented at the annual meetings of the Society for Research on Effective Education, Crystal City, Virginia, March 3-4, 2008.

This research was carried out under funding from the Institute of Education Sciences, U.S. Department of Education (Grant No. R305A040082). However, any opinions expressed are those of the authors and do not necessarily represent IES positions or policies.

Abstract

In systematic reviews of research on educational programs, different reviewers have different policies on measures of content taught in the experimental group but not the control group, called *treatment-inherent* measures. The What Works Clearinghouse (WWC) averages effect sizes from such measures in with those from treatment-independent measures, while the Best Evidence Encyclopedia (BEE) excludes them. This paper contrasts effect sizes from treatment-inherent and treatment-independent measures in WWC and BEE reading and math reviews to explore the degree to which these measures produce different estimates. In all comparisons, treatment-inherent measures produce much larger positive effect sizes than treatment-independent measures. Based on these findings, it is suggested that program effectiveness reviews exclude treatment-inherent measures, although such outcomes may be separately reported.

A key methodological issue in systematic reviews of research on educational programs is the use of measures inherent to skills taught in the experimental group but not the control group in many studies. For example, imagine a study of an innovative approach to biology instruction in which the investigator makes a cogent argument that children should repeat classic experiments from history, such as Pasteur's experiment debunking the principle of spontaneous generation. She assigns a group of classes to do such experiments while another group of classes experiences traditional laboratory experiments in biology.

In this thought experiment, the researcher faces a dilemma. If she makes up a curriculum-specific test composed of questions about spontaneous generation and other historical experiments, which the control group never saw, the experimental group will obviously perform much better. If she gives a traditional survey test of science concepts, it might fail to register important learning from the experimental treatment. She might well give both a test of historical experiments and a survey test, and then argue that if (as is likely) the experimental and control groups do not differ on the survey test but the experimental group scores better on the test of historical experiments, then the experimental group has gained something of value at no cost in terms of traditional learning. This may be a good solution for the individual study, but now imagine that a reviewer is doing a systematic review of research on effective biology programs. Should both the survey test and the curriculum-specific historical experiments test be included in determining achievement effects of various biology programs?

In the thought experiment, the test of historical experiments could be called a test *inherent to the experimental treatment.* That is, although the test assesses knowledge on skills that curriculum experts might deem to be important, the fact that the test's content is not ordinarily taught (and is not taught in this particular control group) means that any additional learning registered on the inherent test cannot be seen as a successful evaluation of the experimental approach, but is only a demonstration that students exposed to the experimental curriculum at least learn something from it. An effect size from a measure of content taught only to the experimental group is really no different from an assertion that the material tested should be important, according to study authors. It does not constitute evidence that the content is in fact valuable beyond itself.

Curricular Reform vs. Instructional Reform

The discussion about how to treat measures inherent to treatments goes to the heart of the difference between curriculum reform and instructional reform. The problem is that when curriculum reformers want to advocate for the teaching or testing of content or skills that are not currently taught, their argument can only rarely be tested in experiments, because it is of little value to simply demonstrate that students taught atypical content score better on a test of that content than students not taught the content. Instead, curriculum reformers must argue for change in terms of international benchmarks, developments in the substantive field (e.g., in science itself, not science education), in technology, or in philosophy. Only occasionally can curricular reformers point to measureable gains on broadly valued outcomes as a result of schools adopting different curricula. For example, a curriculum reformer arguing the value of Latin instruction might point to research showing that studying Latin improves English vocabulary, but testing the Latin students in Latin adds nothing to the argument that Latin should be part of the curriculum.

In contrast, research on instructional methods (such as cooperative learning), holding curriculum constant, has no such problems. Within reason, any widely accepted, valid and reliable test should show the added value of an effective instructional intervention. The evaluation of instructional improvements is perfectly suited to experiments, which ask whether one or another approach is demonstrably better on outcomes of accepted value.

Program evaluation syntheses (such as those of the What Works Clearinghouse and the Best Evidence Encyclopedia) are designed to provide scientifically valid, educationally meaningful summaries of experimental research on various treatments, which means that they are appropriate (in concept) for evaluations of instructional methods but not for evaluations of curricula. The problem is that in practice, educational treatments mix reforms in curriculum and instruction. This is not a problem if one takes the view that current measures (such as standardized tests) are sufficient, if imperfect, measures of what students should know and be able to do. However, it is a huge problem for curriculum reformers, who typically do not accept current widespread measures and argue that there needs to be attention paid to treatments that teach atypical content and measure outcomes on atypical assessments of that content.

These issues lie behind a controversy between the What Works Clearinghouse and a group that carried out a review of research in mathematics for the National Research Council (2004). Although it identified 147 evaluations of 13 National Science Foundation-supported K-12 curricula and 6 commercial curricula, the NRC report decided that "the evaluations were not sufficiently robust to permit confident judgements on individual programs" because none of the evaluations sufficiently met NRC's proposed definition of effectiveness, "an integrated judgement based on interpretation of a number of scientifically valid evaluations that combine social values, empirical evidence, and theoretical rationales" (Confrey, 2006, p. 195). That is, a positive effect size in rigorous experiments was not sufficient; treatments also had to be consistent with "social values and theoretical rationales" determined by experts, not by student achievement data. Along similar lines, Schoenfeld (2006) argued that the What Works Clearinghouse was flawed in focusing on (mostly standardized) test scores that are too narrow, in that they do not test skills that are not widely taught but, he believes, should be taught. He concludes that the WWC should conduct content analyses of the outcome measures used in comparative studies, and place a high value on studies that show gains on outcomes agreed upon by experts to be important. In a response to Schoenfeld's article, Herman, Boruch, Powell, Fleischman, & Maynard (2006) argued that providing content analyses of every measure would be impractical, and would engage the WWC in endless controversy about which outcomes are of the greatest value.

Treatment-Inherent Measures in the What Works Clearinghouse Reviews

The importance of this issue lies in a current debate revolving around the What Works Clearinghouse (WWC) (Slavin, 2008 a, b; Dynarski, 2008). The What Works Clearinghouse includes treatment-inherent measures in its reviews of research on achievement outcomes of educational programs, averaging them in without distinction with outcomes on measures of skills taught equally in experimental and control groups. For example, a series of studies of a phonemic awareness software program called *Daisy Quest* evaluated the experimental program with kindergartners and first graders who were not, according to the authors, being taught phonemic awareness at all (e.g., Barker & Torgesen, 1994, Foster, Erickson, Foster, Brinkman, & Torgesen, 1994). Further, one of the tests of phonemic awareness was a test given on the computer that was closely patterned on the experimental curriculum, which the

control children had of course never seen. Based on these measures, *Daisy Quest* received the highest possible rating ("positive effects") on the What Works Clearinghouse Beginning Reading and Early Childhood Education topic reports, because the studies obtained significantly positive outcomes on these treatment-inherent measures in randomized experiments. Inherent measures of this kind, which typically produce large effect sizes, qualified many programs for "positive effects" ratings in the What Works Clearinghouse. As another example, a study of *Everyday Mathematics* by Carroll (1998) used only an experimenter-made measure of a form of geometry taught in *Everyday Mathematics* but not in control classes, and this single randomized experiment qualified *Everyday Mathematics* for the only "positive effects" rating given in the Middle School Mathematics topic report. A single randomized study of *Saxon Math,* by Williams (1986), used an experimenter-made measure keyed to the experimental program, and even though the very positive effect size found on this measure contradicted the findings of several matched studies, which found near-zero effect sizes on conventional measures of math achievement, *Saxon Math* received a "positive effects" rating on the WWC Elementary Mathematics topic report. An extreme example is in the WWC English Language Learners topic report, where Prater & Bermudez (1993) evaluated a program called *Peer Tutoring and Response Groups.* The outcome measure that qualified this writing program for a "positive effects" rating involved having ELL students write a composition either with the help of their English proficient teammates (in the experimental group) or by themselves (in the control group).

The inclusion of measures inherent to treatments has been defended by WWC leaders on the basis that there is a continuum of alignment between measures and treatments, and it is impossible to draw a clear line between over-aligned (i.e., treatment-inherent) and fair measures (see Herman et al., 2006). In contrast to the WWC position, reviews by Slavin & Lake (in press), Slavin et al. (2007), and Slavin et al. (in press), written as part of the Best Evidence Encyclopedia (www.bestevidence.org), exclude treatment-inherent measures unless curricula in the experimental and control groups are the same. This is a conservative procedure, however, and may exclude measures that are appropriately sensitive to experimental treatments yet still fair to control groups.

In order to illuminate the consequences of including or excluding treatment-inherent measures, the present paper examined studies included in the What Works Clearinghouse and Best Evidence Encyclopedia (BEE) reviews that used treatment-inherent as well as treatment-independent (usually standardized) measures of achievement to learn how much difference these measures make in effect size estimates and to see if there are well-defined circumstances in which experimenter-made measures may be acceptable in program evaluation syntheses.

Methods

The data for the present study were obtained from studies accepted for inclusion in the What Works Clearinghouse beginning reading, elementary mathematics, and middle school mathematics topic reports (What Works Clearinghouse, 2008a, b, c) and studies included in the elementary and secondary mathematics reviews in the Best Evidence Encyclopedia (Slavin & Lake, in press; Slavin et al., 2007). In each case, studies were included in Tables 1-3 if they used at least one measure deemed to be a treatment-inherent measure made by the experimenter or by the publisher of the curriculum. Effect sizes from the WWC or BEE reviews are then averaged across studies for treatment-inherent as well as treatment-independent measures.

Results

Table I summarizes the results from the seven studies that used treatment-inherent tests, accepted by the What Works Clearinghouse in its elementary and middle school mathematics topic reports. Two of the studies used only treatment-inherent tests, and five used both treatment-independent and treatment-inherent tests.

As the Table makes clear, effect sizes on treatment-inherent tests are consistently and substantially higher than those found on treatment-independent tests. The overall mean was +0.45 for tests inherent to treatment but -0.03 on independent tests. Within studies, differences were marked; in three of the five What Works Clearinghouse math studies that used both treatment-inherent and treatment-independent measures, effect sizes for the two types of measures were in opposite directions.

Table 1: Comparison of Effect Sizes for Mathematics Studies with Treatment-Inherent and Treatment-Independent Measures: What Works Clearinghouse

Study	Program	Measures	Effect Sizes	
			Treatment-Inherent	Treatment-Independent
Carroll (1998)	Everyday Mathematics	Researcher-developed geometry test	+0.37	
Ridgeway et al (2002)	Connected Mathematics	ITBS		-0.20
		Balanced assessment test	+0.27	
Williams (1986)	Saxon Math	End-of-course test	+0.65	
Peters (1992)	UCSMP	Orleans-Hanna		-0.13
		Understanding of algebraic components	+0.28	
Hedges et al (1986)	Transition Mathematics (UCSMP)	Orleans-Hanna		+0.17
		HSST: General math		+0.13
		Geometry readiness	+0.29	
Thompson et al (2005)	Transition Mathematics (UCSMP)	HSST: General math		-0.26
		Algebra readiness	+0.09	
		Geometry readiness	+0.51	
		Problem solving and understanding	+0.35	
Thompson et al (2005)	UCSMP Algebra	HSST: Algebra		+0.12
		Algebra readiness	+0.78	
		Problem solving and understanding	+0.89	
Mean			+0.45	-0.03

Table 2

Comparison of Effect Sizes for Mathematics Studies with Treatment-Inherent and Treatment-Independent Measures: Best Evidence Encyclopedia

Study	Program	Measures	Effect Sizes	
			Treatment-Inherent	Treatment-Independent
Snider & Crawford (1996)	CMC	NAT		+0.26
		CMC Test	+3.49	
		Facts Fluency	+2.03	
Crawford & Snider (2000)	CMC	NAT		+0.41
		CMC Test	+3.25	
		Facts Fluency	+0.95	
Ysseldyke et al., (2006)	Accelerated Math	NALT		+0.19
		STAR	+0.35	
Ysseldyke et al., (2003)	Accelerated Math	NALT		+0.08
		STAR	+0.20	
Median			+1.49	+0.23

Note: The Best-Evidence Encyclopedia excludes findings from treatment-inherent measures

Table 2 shows effect sizes from four additional mathematics studies identified by the Best Evidence Encyclopedia reviews but not the What Works Clearinghouse (which only reviewed commercial textbooks). The BEE does not incorporate effect sizes from treatment-inherent measures, but the Table adds effect sizes from treatment-inherent measures to the BEE effect sizes from treatment-independent measures. Two of the studies, by Snider & Crawford (1996) and Crawford & Snider (2000), evaluated a program developed at the University of Oregon called Connecting Mathematics Concepts, or CMC. In both studies, multiple measures were used. A CMC-specific measure produced huge effect sizes (+3.49 and +3.25). Another CMC-constructed test, facts fluency, also produced substantial effect sizes (+2.03 and +0.95). In contrast, effect sizes on a standardized National Achievement Test had effect sizes of +0.41 and +0.26. The authors also used a test aligned with the Scott Foresman control group, and found effect sizes strongly supporting CMC (+0.96 and +1.25), but these are of course smaller than those for the CMC tests.

Because of the exceptionally large means in the CMC studies, medians rather than means were used to summarize the effect sizes in Table 2. These were +1.49 for treatment-inherent tests and +0.23 for treatment-independent tests.

Table 3 shows effect sizes for ten studies from the What Works Clearinghouse beginning reading topic report. Once again, treatment-inherent measures were associated with far more positive effect sizes (mean ES=+0.51) than were treatment-independent measures (mean ES=+0.06).

One of the *Daisy Quest* studies, by Mitchell & Fox (2001), illustrated an important aspect of the issue of treatment-inherent measures. This study had three treatment groups. In one, K-1 students experienced the *Daisy Quest* phonemic awareness software. In a control treatment, children used math and drawing software. In a third group, teachers taught the same content as that emphasized in *Daisy Quest*, but children did not use computers. The outcome measures were specific to the *Daisy Quest* content. In the comparison of *Daisy Quest* to control, the curriculum specific measures were considered treatment-inherent, because the control group was not receiving the same content. However, in the comparison between *Daisy Quest* and the teacher, both groups were studying the same content, so the same measures were considered treatment-independent. As the Table shows, the outcomes from the treatment-inherent and treatment-independent comparisons were diametrically opposed (+0.85 vs. -0.46).

Table 3

Comparison of Effect Sizes for Beginning Reading Studies with Treatment-Inherent and Treatment-Independent Measures: What Works Clearinghouse

Study	Program	Measures	Effect Sizes	
			Treatment-Inherent	Treatment-Independent
Ross et al (2004)	Accelerated Reader	STAR Reading	+0.31	
		STAR Early Literacy	+0.43	
Barker & Torgerson (1995) (means of two comparisons)	Daisy Quest	Phonological awareness (5 measures)	+0.70	
		Phonics (4 measures)		+0.30
Foster et al (1995) (means of two comparisons)	Daisy Quest	Phonological awareness (4 measures)	+0.90	
Mitchell & Fox (2001)	Daisy Quest	Phonological awareness (4 measures, compared to untreated)	+0.85	
		Phonological awareness (4 measures, compared to teacher instruction)		-0.46
Taylor et al (1991)	Early Intervention in Reading	Gates-MacGinitie		+0.47
		Segmentation & blending	+0.80	
		Vowel sounds	+1.39	

Mathes & Babyak (2001)	PALS	Oral reading fluency	+0.51	
		Phonological awareness	+0.69	
Mathes et al (1998)	PALS	Oral reading fluency	+0.37	
Mathes et al (2003) (mean of two comparisons)	PALS	Woodcock Word ID		+0.15
		Woodcock Passage Comp.		-0.10
		Oral reading fluency	+0.13	
Hancock (2002)	Read Naturally	Peabody Picture Vocabulary Test		+0.02
		Oral reading fluency	+0.16	
		Word use fluency	+0.22	
		CBM: Cloze	-0.08	
Mesa (2004)	Read Naturally	Oral reading fluency	+0.23	
Mean			+0.51	+0.06

Conclusion

The data summarized in Tables 1-3 demonstrate the powerful impact of measures inherent to treatment on estimates of effect sizes in program effectiveness reviews. In every case, effect sizes for measures inherent to treatments were very positive (+0.45, +0.51, and +1.49), while those for measures independent of treatments were mostly near zero (-0.03, +0.06, and +0.23). Comparisons within studies consistently found more positive effects for inherent measures.

As noted earlier, what is disturbing about these findings is that the What Works Clearinghouse averages effect sizes from treatment-inherent measures into its effect size estimates without comment. Frequently, program ratings made by the WWC depend entirely or mostly on findings from measures inherent to treatments. The WWC places a strong emphasis on random assignment, which is appropriate, but in ignoring issues such as the use of treatment-inherent measures it makes final ratings that do not correspond to common sense. If the ratings in the What Works Clearinghouse or similar reviews were to become important to users or producers of educational programs, it would be easy to imagine that program developers or advocates would increasingly carry out or commission studies using only (or primarily) treatment-inherent measures, knowing that these are certain to produce large positive effects.

It is important to note that the degree to which treatment-inherent and treatment-independent measures produce different outcomes in the same studies varies considerably. For example, studies of *Accelerated Math* by Ysseldyke et al. (2003, 2006) used a STAR measure produced by the publisher of *Accelerated Math* but not, apparently, excessively aligned with it, and though effect sizes for STAR tests are always larger than for those of the standardized (and independent) NALT, the differences are modest. Similarly, studies of PALS (e.g., Mathes et al., 2003) find modest differences between effect sizes for measures created by the authors of PALS and those of independent measures. In contrast, substantial differences exist between treatment-inherent and treatment-independent measures for most other programs. The different patterns might justify trying to find a way of determining how much a given set of curriculum-specific measures differs from a treatment-independent test. However, there is no clear way to do this a priori. As a result, the only practical solution is to exclude measures that are inherent to treatments in program effectiveness reviews. Most studies that use treatment-inherent measures also use treatment-independent measures, so excluding inherent measures does not entirely exclude very many studies.

In research on curricular innovations, there is nothing wrong in using treatment-inherent measures as part of a formative evaluation process, perhaps leading over time to evaluations on treatment-independent measures (see Clements, 2007). However, as the present findings make clear, mixing up effect sizes from treatment-inherent and treatment-independent measures makes no sense. Readers need to know that effect sizes averaged across studies can be interpreted in a consistent way, as indications of improved performance on measures of content taught in all conditions.

If evidence-based reform is to prevail in educational practice, educators must have meaningful, scientifically-valid reviews of research to use in deciding which programs and practices are truly supported by strong research. Educators have a right to know how various programs they might implement are likely to help their students improve their

achievement on standard, widely understood measures. Results from measures of content not taught in the control group may be of interest to some educators and curriculum reformers, but at a minimum such measures must be separately identified and discussed. Fair evidence from measures that were fair to the control group is the most defensible basis for evidence-based policies and practices.

References

Barker, T., & Torgesen, J. K. (1995). An evaluation of computer-assisted instruction in phonological awareness with below average readers. *Journal of Educational Computing Research, 13* (1), 89–103.

Carroll, W. M., & Fuson, K. C. (1998). *A comparison of Everyday Math (EM) and McMillan (MC) on Evanston student performance on whole-class tests: Recommendations for revision of Everyday Mathematics Grades 1, 2, 3, and 4.* (Available from Karen C. Fuson, School of Education and Social Policy, Northwestern University, 2115 N. Campus Drive, Evanston, IL 60208-2610).

Clements, D. H. (2007). Curriculum research: Toward a framework for "research-based curricula." *Journal for Research in Mathematics Education, 38* (1), 35-70.

Confrey, J. (2006). Comparing and contrasting the National Research Council report *On Evaluating Curricular Effectiveness* with the What Works Clearinghouse approach. *Educational Evaluation and Policy Analysis, 28* (3), 195-213.

Crawford, D.B. & Snider, V.E. (2000). Effective mathematics instruction: The importance of curriculum. *Education and Treatment of Children, 23*(2), 122-142.

Dynarski, M. (2008). Bringing answers to educators: Guiding principles for research syntheses. *Educational Researcher, 37* (1), 27-29.

Foster, K. C., Erickson, G.C., Foster, D.F., Brinkman, D. & Torgeson, J.K. (1994). Computer administered instruction in phonological awareness: Evaluation of the DaisyQuest program. *Journal of Research and Development in Education, 27*, 126-137.

Foster, K. C., Erickson, G. C., Foster, D. F., Brinkman, D., & Torgesen, J. K. (1994). Computer administered instruction in phonological awareness: Evaluation of the *DaisyQuest* program. *Journal of Research and Development in Education, 27* (2), 126–137.

Hancock, C. M. (2002). Accelerating reading trajectories: The effects of dynamic research-based instruction. *Dissertation Abstracts International, 63* (06), 2139A. (UMI No. 3055690)

Hedges, L. V., Stodolsky, S. S., Mathison, S., & Flores, P. V. (1986). *Transition Mathematics Field Study*. Chicago, IL: University of Chicago School Mathematics Project.

Herman, R., Boruch, R., Powell, R., Fleischman, S., & Maynard, R. (2006). Overcoming the challenges: A response to A. Schoenfeld's "What Doesn't Work". *Educational Researcher, 35* (2), 22-23.

Mathes, P. G., & Babyak, A. E. (2001). The effects of peer-assisted literacy strategies for first-grade readers with and without additional mini-skills lessons. *Learning Disabilities Research & Practice*, *16* (1), 28–44.

Mathes, P. G., Howard, J. K., Allen, S. H., & Fuchs, D. (1998). Peer-assisted learning strategies for first-grade readers: Responding to the needs of diverse learners. *Reading Research Quarterly*, *33* (1), 62–94.

Mathes, P. G., Torgesen, J. K., Clancy-Menchetti, J., Santi, K., Nicholas, K., Robinson, C., et al. (2003). A comparison of teacher-directed versus peer-assisted instruction to struggling first-grade readers. *The Elementary School Journal*, *103* (5), 459–479.

Mesa, C. L. (2004). *Effect of Read Naturally software on reading fluency and comprehension*. Unpublished master's thesis, Piedmont College, Demorest, GA.

Mitchell, M.J. & Fox, B. J. (2001). The effects of computer software for developing phonological awareness in low-progress readers. *Reading Research and Instruction, 40* (4), 315-332.

National Research Council (2004). *On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations*. Washington, DC: National Academies Press.

Peters, K. G. (1992). Skill performance comparability of two algebra programs on an eighth-grade population. *Dissertation Abstracts International, 54(01),* 77A. (UMI No. 9314428).

Prater, D. L., & Bermudez, A. B. (1983). Using peer response groups with limited English proficient writers. *Bilingual Research Journal, 17* (1, 2), 99–116.

Ridgway, J. E., Zawojewski, J. S., Hoover, M. N., & Lambdin, D. V. (2002). Student attainment in the Connected Mathematics curriculum. In S. L. Senk & D. R. Thompson (Eds.), *Standards-based school mathematics curricula: What are they? What do students learn?* (pp. 193-224). Mahwah, NJ: Lawrence Erlbaum Associates, Inc

Ross, S. M., Nunnery, J., & Goldfeder, E. (2004). *A randomized experiment on the effects of Accelerated Reader/Reading Renaissance in an urban school district: Preliminary evaluation report.* Memphis, TN: The University of Memphis, Center for Research in Educational Policy.

Schoenfeld, A. (2006). What doesn't work: The challenge and failure of the What Works Clearinghouse to conduct meaningful reviews of studies of mathematics curricula. *Educational Researcher, 35* (2), 13-21.

Snider, V.E., & Crawford, D.B. (1996). Action research: Implementing Connecting Math Concepts. *Effective School practices, 15* (2), 17-26.

Slavin, R. E. (2008a). What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37* (1), 5-14.

Slavin, R.E. (2008b). Evidence-based reform in education: Which evidence counts? *Educational Researcher, 37* (1), 47-50.

Slavin, R.E., & Lake, C. (in press). Effective programs in elementary math: A best evidence synthesis. *Review of Educational Research.*

Slavin, R. E., Lake, C., & Groff, C. (2007). *Effective programs in middle and high school mathematics: A best-evidence synthesis*. Manuscript submitted for publication.

Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (in press). Effective reading programs for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly*.

Taylor, B. M., Frye, B. J., Short, R., & Shearer, B. (1991). *Early Intervention in Reading: Preventing reading failure among low-achieving first grade students.* Minneapolis: University of Minnesota, Center for Urban and Regional Affairs and Office of the Vice President of Academic Affairs.

Thompson, D.R., Senk, S.L., Witonsky, D., Usiskin, Z., Kaeley, G. (2005). *An evaluation of the second edition of UCSMP Transition Mathematics.* Chicago, IL: University of Chicago School Mathematics Project.

What Works Clearinghouse (2008a). Beginning reading. What Works Clearinghouse Topic Report. At www.whatworks.ed.gov

What Works Clearinghouse (2008b). Elementary school mathematics. What Works Clearinghouse Topic Report. At www.whatworks.ed.gov.

What Works Clearinghouse (2008c). Middle school math. What Works Clearinghouse Topic Report. At www.whatworks.ed.gov.

Williams, D.D. (1986). *The incremental method of teaching Algebra 1.* Research report, University of Missouri-Kansas City.

Ysseldyke, J., Spicuzza, R., Kosciolek, S., & Boys, C. (2003). Effects of a learning information system on mathematics achievement and classroom structure. *Journal of Educational Research, 96* (3), 163-173.

Ysseldyke, J.E. & Bolt, D. (2006). *Effect of technology-enhanced progress monitoring on math achievement.* Minneapolis, MN: University of Minnesota.