Running Head: Measures Inherent to Treatments

Measures Inherent to Treatments in Program Effectiveness Reviews

Robert E. Slavin

Nancy A. Madden

Johns Hopkins University

-and-

University of York

200 W. Towsontown Boulevard
Baltimore, MD 21204
T: 410-616-2310
F: 410-324-4440
rslavin@jhu.edu
nmadden@jhu.edu

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.

Measures Inherent to Treatments in Program Effectiveness Reviews

Abstract

Program effectiveness reviews in education seek to provide educators with scientifically valid and useful summaries of evidence on achievement effects of various interventions. Different reviewers have different policies on measures of content taught in the experimental group but not the control group, called here *treatment-inherent* measures. These are contrasted with *treatment-independent* measures of content emphasized equally in experimental and control groups. The What Works Clearinghouse (WWC) averages effect sizes from such measures with those from treatment-independent measures, while the Best Evidence Encyclopedia (BEE) excludes treatment-inherent measures. This article contrasts effect sizes from treatment-inherent and treatment-independent measures in WWC 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, or at least report them separately.

In recent years, there has been a rapid development of program effectiveness reviews, systematic reviews of research on the outcomes of educational programs. The federally supported What Works Clearinghouse (WWC) is carrying out reviews of research to attempt to identify programs with scientifically valid evidence of effectiveness. Other review series, such as the Best Evidence Encyclopedia (BEE), the EPPI Centre in the UK, and Social Programs that Work, have created web sites for educators as well as researchers to give them fair, consistent, and useful summaries of the strength of evidence supporting various practical programs. These reviews are important in themselves, but they are also key linchpins in the larger movement toward evidence-based reform. Educators and policy makers are not likely to use programs with strong evidence of effectiveness unless there are clear and trustworthy sources of information on what works.

All series of program effectiveness reviews seek to apply consistent review standards and procedures to synthesize research in a given area. However, different review series use markedly different standards and procedures, and therefore come to different conclusions. There are many issues on which different review series disagree, such as standards for study inclusion, how to compute effect sizes, how much to emphasize random assignment, how to deal with study sample sizes, and how to pool findings from individual studies to come up with overall ratings.

A 2008 special issue of the *Educational Researcher* contained an article by Slavin (2008a) and several responses by distinguished scholars on these important methodological issues. Clearly, the field is at an early stage in understanding all of the key decisions reviewers need to take into account. The rapid development of program effectiveness reviews in education and the interest these reviews have drawn from government as well as research and practitioner audiences make it imperative that methodologies for systematic reviews of program evaluations in education advance as rapidly as possible.

Treatment-Inherent Measures in Program Effectiveness Reviews

One of the key methodological issues discussed by Slavin (2008a) involves how to deal with measures of content taught in experimental but not control groups. 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, about which the control group was never taught, 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, and may be useful for many theoretical and practical purposes, but now imagine that a reviewer is doing a systematic review of research on effective biology programs. Should effect sizes from the survey test and the curriculum-specific historical experiments test simply be averaged 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 or 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 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, or that the method assessed is an effective way to teach that content. The opposite of a treatment-inherent measure could be called a *treatment-independent measure*, one that measures skills or content taught in the control as well as the experimental group.

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 (and showing that these students learned much more Latin than a control group that did not study Latin) adds nothing to the argument that Latin should be part of the curriculum. In contrast, a test of English vocabulary would be

fair to both groups. Imagine that in a study of Latin teaching (compared to no Latin teaching) there was an effect size of +2.0 on a Latin measure and 0.00 on an English vocabulary measure. In a synthesis of vocabulary interventions, these different findings should not of course be averaged. The Latin results might be reported in the original study, but they have no place in the synthesis.

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 achievement test should show the added value of an effective instructional intervention. The evaluation of instructional processes is perfectly suited to experiments, which ask whether one or another approach is demonstrably better on outcomes of accepted value.

Program evaluation reviews (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. They are intended to give practicing educators useful information about the likely impacts on student achievement of implementing various educational programs. Their intention is not primarily to advance theory, but to provide educators with the kind of information that "Consumer Reports" provides on cars or refrigerators. In such program evaluation syntheses, the intended audience wants to know how to improve achievement in mathematics or reading, for example, and atypical measures of skills or content not taught in the control group are of little interest.

Treatment-Inherent Measures in the What Works Clearinghouse Reviews

The importance of the question of how to handle treatment-inherent measures lies in a current debate revolving around the What Works Clearinghouse (Slavin, 2008 a, b; Dynarski, 2008). Although the What Works Clearinghouse theoretically excludes measures that are

"overaligned" with the treatment, in practice it 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 or content 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 in comparison to control students who were not, according to the authors, being taught phonemic awareness at all (e.g., Barker & Torgesen, 1995, 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 activities that were central to the experimental curriculum, which the control children had of course never seen. In one *Daisy Quest* study, by Mitchell & Fox (2001), there was a control condition in which a teacher (instead of the computer) taught the phonemic awareness skills. The same measures considered treatment-inherent in comparison to children not taught phonemic awareness could be considered treatment-independent in comparison with the same content taught by a teacher. Based on an average of many treatment-inherent and a few treatment-independent measures, *Daisy Quest* received the highest possible rating ("positive effects") on the What Works Clearinghouse (2008a) Beginning Reading topic report and the WWC (2008b) Early Childhood Education topic report, because the studies obtained significantly positive outcomes on these (mostly) treatment-inherent measures in randomized experiments.

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 "potentially positive effects" rating given in the WWC (2008c) Elementary 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 one randomized and several matched studies, which found near-zero effect sizes on conventional measures of math achievement, *Saxon Math* was one of only two programs that received a "positive effects" rating on the WWC (2008d) Middle School Mathematics topic report.

Measures inherent to treatments are usually ones that were made by the researchers who carried out the evaluation or by the publishers of the curriculum. For example, studies of *Accelerated Reader* and *Accelerated Math* (e.g., Ysseldyke et al., 2003; Ysseldyke & Bolt, 2006) routinely use a computerized test called STAR that is also used in these programs as a progress check. However, standardized tests can also be treatment-inherent measures when the skills they assess have not yet been taught to the control group. For example, imagine a program that taught algebra to sixth graders. It could use a standardized algebra test, but this would still be inherent to the treatment if the control classes were not receiving instruction in algebra. A follow-up test given after the control group had also received instruction in algebra might fairly assess the outcomes of early introduction of algebra, but a test given before the control group has received any algebra instruction would be treatment-inherent.

In practice, this type of treatment-inherent measurement (early assessment of skills taught earlier than usual) is primarily seen in studies of kindergartners, when phonological awareness skills usually taught in first grade are taught early to an experimental group but not a control group. In these studies, the control kindergartners were not being taught any phonological awareness, or were not being taught to read at all, so effect sizes on end-of-kindergarten phonological awareness measures are invariably substantial. Such studies (e.g., Lundberg, Frost, & Petersen, 1988; Blachman et al., 1999) typically follow up children into first and second grades to determine the lasting effects of early introduction of phonological awareness training. The follow-up reading measures are treatment independent, because by the end of first grade all children have been taught to read. However, the end-of-kindergarten

measures, which are typically minimized in importance by the authors themselves, are treatment-inherent.

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 treatment-independent measures (Whitehurst, personal communication, 2006; Herman et al., 2006). In contrast to the WWC position, reviews by Slavin & Lake (2008), Slavin, Lake, & Groff (in press), and Slavin, Cheung, Groff, & Lake (2008), written as part of the Best Evidence Encyclopedia (www.bestevidence.org), exclude treatment-inherent measures.

The main argument made by Herman et al. (2006), Whitehurst (2006), and others against separating treatment-inherent and treatment-independent measures is that this division cannot be made reliably. Yet in practice this is usually not difficult. The Best Evidence Encyclopedia, for example, uses the following decision rules.

1. If a skill or concept has been taught to the experimental group but not to the control group, all measures of that skill or concept are treatment-inherent.

2. Except for #1, standardized tests and other assessments not made by the developer or the experimenter are treatment-independent.

3. Experimenter-made or developer-made assessments are treatment-independent if curriculum is held constant in experimental and control groups (i.e., the groups differ in teaching method but not content). Otherwise, such assessments are treatment-inherent.

4. If an experimental treatment gives students extensive practice with an unusual response format (such as computer-adaptive testing) to which the control group is not exposed, then measures using this response format are treatment-inherent.

Application of these decision rules is usually straightforward. There are sometimes difficult decisions in individual studies that have to be discussed among independent reviewers, but this is true of many distinctions in systematic reviews.

In order to illuminate the consequences of including or excluding treatment-inherent measures, the present article examined studies included in three What Works Clearinghouse 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.

Methods

The data for the present study were obtained from studies accepted for inclusion in the What Works Clearinghouse beginning reading, elementary school mathematics, and middle school mathematics topic reports as of February, 2008 (What Works Clearinghouse, 2008a, c, d). In each case, studies were included in Tables 1 and 2 if they used at least one measure deemed to be a treatment-inherent measure. Treatment-inherent and treatment-independent measures were defined using the Best Evidence Encyclopedia decision rules described above. All studies accepted by the WWC that were deemed to have used inherent measures in reading or math were included in Table 1 or 2. Effect sizes from the WWC reviews were then averaged across studies for treatment-inherent as well as treatment-independent measures. The WWC (2008e) computes effect sizes as the difference between experimental and control

means divided by the pooled within-group standard deviation, with adjustments for clustering, and then averages unweighted effect sizes. We used the effect sizes computed by the WWC and averaged them in the same way.

======

TABLE 1 HERE

======

Results

Table 1 summarizes the results from the seven studies that used treatment-inherent tests, accepted by the What Works Clearinghouse (2008c, d) 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. Restricting attention to the five studies that used both treatment-inherent and treatment-independent measures, the mean effect sizes were +0.43 for treatment-inherent and -0.03 for treatment-independent measures.

======

TABLE 2 HERE

=======

Table 2 shows effect sizes for the ten studies from the What Works Clearinghouse beginning reading topic report that used treatment-inherent measures. 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). Across five studies that reported effect sizes for both types of measures, mean effect sizes were +0.52 for treatment-inherent measures and +0.06 for treatment-independent measures.

One of the *Daisy Quest* studies, by Mitchell & Fox (2001), mentioned earlier, 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 unrelated to reading, but received no instruction in phonemic awareness at any time. In a third group, teachers taught the same phonemic awareness 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 any instruction in phonemic awareness. However, in the comparison between *Daisy Quest* and the teacher, both groups were receiving phonemic awareness training, 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).

Conclusion

The data summarized in Tables 1 and 2 demonstrate that effect sizes on measures inherent to treatment are very different from those on measures independent of treatments. In every case, effect sizes for measures inherent to treatments were very positive (+0.45 and

+0.51) while those for measures independent of treatments were mostly near zero (-0.03 and +0.06). Comparisons within studies consistently found more positive effects for treatment-inherent measures than for treatment-independent measures.

As noted earlier, the importance of these findings is in calling into question the practices of the What Works Clearinghouse (2008e), which averages effect sizes from treatment-inherent measures into its effect size estimates without distinction. Frequently, positive program ratings made by the WWC depend entirely or mostly on inflated findings from measures inherent to treatments. 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 likely to produce large positive effects.

It is important to note that for many purposes, effect sizes from treatment-inherent measures may be of value. For example, such measures may be useful in theory building. 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). In individual studies, it may be appropriate to report findings on treatment-inherent measures. The problem of how to handle effects on treatment-inherent measures arises only when synthesizing findings, as in program effectiveness reviews. As the present findings make clear, averaging effect sizes from treatment-inherent and treatment-independent measures cannot be justified in program effectiveness reviews intended to give educators unbiased information on the likely achievement outcomes of various interventions. 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 truly have strong evidence of positive effects on valid measures of common achievement objectives. Results from measures of content not taught in the control group may be useful in some types of research and may be of interest to some educators and curriculum reformers, but at a minimum such measures must be separately identified and discussed. Evidence from measures that were fair to the control group is the most meaningful and 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.

Blachman, B.A., Tangel, D., Ball, E., Black, R., & McGraw, C. (1999). Developing phonological awareness and word-recognition skills: A two-year intervention with low-income inner-city children. *Reading and Writing: An Interdisciplinary Journal, 11*, 239-273.

Carroll, W. M. (1998). Geometric knowledge of middle school students in a reform-based mathematics curriculum. *School Science and Mathematics, 98* (4), 188-197.

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

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., & 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.

Lundberg, I., Frost, J., & Peterson, O. (1988). Effects of an extensive program for stimulating phonological awareness in pre-school children. *Reading Research Quarterly, 23*, 263-284.

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.

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).

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.

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. (2008). Effective programs in elementary math: A best evidence synthesis. *Review of Educational Research, 78* (3), 427-515

Slavin, R. E., Lake, C., & Groff, C. (in press). Effective programs in middle and high school mathematics: A best-evidence synthesis. *Review of Educational Research*.

Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (2008). Effective reading programs for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly, 43* (3), 290-322.

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.

Thompson, D. R., Senk, S. L., Witonsky, D., Usiskin, Z., & Kaely, G. (2006). *An evaluation of the second edition of UCSMP Algebra.* Chicago, IL: University of Chicago School Mathematics Project.

What Works Clearinghouse (2008a). *Beginning reading*. What Works Clearinghouse Topic Report. At www.ies.ed.gov/ncee/wwc.

What Works Clearinghouse (2008b). *Early childhood education.* What Works Clearinghouse Topic Report. At www.ies.ed.gov/ncee/wwc.

What Works Clearinghouse (2008c). *Middle school math*. What Works Clearinghouse Topic Report. At www.ies.ed.gov/ncee/wwc.

What Works Clearinghouse (2008d). *Elementary school mathematics*. What Works Clearinghouse Topic Report. At www.ies.ed.gov/ncee/wwc.

What Works Clearinghouse (2008e). *WWC procedures and standards handbook.* (version 2.0). At www.ies.ed.gov.ncee/wwc.

Whitehurst, R. (2006, October 31). Personal communication.

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.

Table 1

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

Study	Program	Measures	Effect Sizes	
			Treatment-Inherent	Treatment-Independent
Carroll & Fuson (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 Algebra	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., 2006	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 What Works Clearinghouse Effect Sizes for Beginning Reading Studies with Treatment-Inherent and Treatment-Independent Measures

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