2010

A Comparison of Results Obtained Using Standardized Tests and Researcher-Developed Tests in Intervention Studies.

Barak Rosenshine

University of Illinois at Urbana-Champaign

One question that interests those conducting and reading the results of experimental classroom-based studies is whether different measures of achievement yield different results. Typically, in studies of cooperative learning, for example, the same content is taught to students in both the experimental group and the control group. At the end of the instructional period, all students receive a test. In some studies, the test was developed by the experimenter; but in other studies, a standardized test was used. Do the results vary according to which type of test was given? Were different results obtained when standardized tests or experimenter-developed tests were used to compare the achievement of students in experimental and control groups?

I attempted to study this question by locating meta-analyses that contained studies in which both a researcher-developed test or a standardized-test was used as an outcome measure, and where the reviewer reported separate results for each of the two types of outcome measures. For example, Bredderman, 1983) reported on studies in which activity-based learning was the treatment. He located 30 studies in which researcher-developed tests were used and 22 studies where standardized-tests were used. His results and the results of another 10 similar meta-analyses are presented in Table 1. For each study in Table 1 the first one cell gives the results, expressed in average effect sizes, for

studies where researcher-developed tests were used to assess student achievement and the other cell gives the results for studies where standardized-tests were used. Thus, in Bredderman (1983) review, the average effect size was higher when researcher-developed test were given.

The search was limited to meta-analyses where the authors that reported separate summary results for studies that used standardized tests and for studies that used researcher-developed tests. Meta-analyses where all test results were lumped into a single "achievement" category as was done by Hattie, Biggs, and Purdie (1996) and by Wang, Haertel, and Walberg (1993) were not included.

Effect Sizes.

Eleven, non-overlapping, meta-analysis of intervention studies in a number of subject areas and using a number of instructional formats were found. The authors of these reviews reported the 'effect sizes" of the results. An effect size is a measure of the size of the difference in test scores between the experimental and the control group. An effect size is expressed in standard deviation units, and is typically computed by subtracting the control group mean score from the experimental group mean score and dividing the result by the standard deviation of the control group or the standard deviation of both experimental and control groups.

Effect sizes can be converted into percentiles using a table of the normal curve. For example, an effect size of. .33 means that someone at the 50th percentile of the control group would have been at the 63rd percentile of the control group had they received the experimental treatment. An effect size of .15 means that someone at the 50th

percentile of the control group would have scored at the 56th percentile of the control group had they received the experimental treatment.

Summary of meta-analyses.

The 11 meta-analyses that were found summarized the results of intervention studies in elementary science, mathematics, reading comprehension, general science, social studies, biology, physics, chemistry, and geography. Students ranged from elementary school through college, and two studies focused on students with disabilities. Seven of the meta-analyses summarized the results of instruction in specific instructional arrangements. These arrangements were:

Activity-based learning (Bredderman, 1983),

Ability grouping (Kulik et al., 1982),

Small group learning (Lou at al., 1996),

Keller's Personalized System of Instruction (Kulik, 1979),

Mastery learning (Slavin, 1987; Kulik et al., 1990) and

Cooperative learning (Slavin, 1990).

A variety of subjects were taught in studies that employed each of the six instructional formats listed above. For example, Kulik et al., (1990) wrote that math, biology, physics, chemistry, general science, education, psychology, social studies, reading, geology, or ecology were taught in their review of studies that used the Personalized System of Instruction or the Mastery Learning format.

There were also four meta-analyses where only a single subject was taught in the studies that were assembled. Bredderman (1983)'s review only contained studies in

elementary science, Gersten et al., (2009) in mathematics for students with disabilities, Rosenshine and Meister (1994) and Rosenshine et al., (1996) only assembled studies where reading comprehension was the outcome measure.

To illustrate the reviews, here is a summary of the results in five of the mete-analyses: activity based science, Mastery Learning, Cooperative Learning, and cognitive strategy instruction in summarization and generating questions.

Activity-based science. The earliest meta-analysis relevant to this issue that appeared in the Review of Educational Research (RER) was by Bredderman (1982). Bredderman presented the results for 57 controlled studies which used one of three activity-based science programs: Elementary School Science, Science --A Process Approach, and The Science Curriculum Improvement Study, The results, in Table 1, show a mean effect size of .25 when standardized tests were used and an overall effect size of .38 when researcher-developed tests were used.

Mastery Learning. There were two reviews of the research on Mastery Learning in RER: one by Slavin (1987), followed by one by Kulik, Kulik, and Brangert-Downs (1990). Slavin only studied Mastery Learning in elementary and secondary schools. In one part of his review, Slaivin looked at Mastery Learning studies where equal time was provided for experimental and control groups. Slavin obtained a median effect size of .04 for the seven studies that used standardized tests and a median effect size of .27 for the nine studies that used researcher-developed tests (Table 1). The standardized tests were in math and reading. The researcher-developed tests were in reading, math, anthropology, Spanish, chemistry and algebra.

Kulik, Kulik, and Bangert-Drowns (1990) reported on 36 studies that used a mastery learning approach. Five studies used standardized tests, and these had an average effect size of .07. The effect size for the 31 studies that used researcher-developed tests was .65. Standardized tests were used in math and reading, and researcher-developed tests were used in math, physics, science, social science, chemistry, psychology, and geometry.

When J.Kulik, Kulik and Bangert-Drowns (1990) conducted an analysis of only those Mastery Learning studies which they and Slavin had in common, the results were an effect size of .09 for the four studies where standardized tests were used and .36 or .45 for the eight common studies where researcher-developed tests were used (Table 1). Kulik et al. (1990) also summarized the results for 72 college studies that used Keller's Personalized System of Instruction, (PSI) a computer-based form of Mastery Learning. The mean effect size for the nine PSI studies was that used standardized tests were .30 and the mean effect size for the 57 studies that used researcher-developed tests was .52. The standardized tests were in economics, psychology, psychiatry, chemistry, biochemistry, and math.

Cooperative learning and small-group learning. Lou, Abrami, Spence, Poulsen, Chambers, and d'Apollonia (1966) reported on the results of 66 studies that used different forms of within-class grouping. The mean effect size for the 18 studies that used standardized tests was .07. The mean effect size was .34 when researcher-developed tests were used and .42 when teacher-made tests were used. The standardized tests were mathematics, physical science, and chemistry.

Cognitive strategy instruction. Rosenshine and Meister (1994, 1996) reviewed the research on cognitive strategy instruction in reading. Typically, in these studies, students in the experimental group are taught one or more cognitive strategies during the time scheduled for reading while students in the control group continue with their regular reading instruction. Teaching students to generate questions based on the material they are reading is one such cognitive strategy. Teaching students to summarize a short passage is another. At the end of the instruction or at the end of the semester, students in some studies took a researcher-developed test and students in other studies took a standardized test in reading. In both types of tests, students read new passages and answered multiple-choice questions about the passages.

Rosenshine and Meister (1994, 1996) assembled three separate sets of studies. The first sample (Rosenshine and Meister, 1994) consisted of 24 studies in which students were taught the cognitive strategy of summarization. The second sample (Rosenshine and Meister, 1996) consisted of 20 studies in which students were taught the cognitive strategy of question-generation. The third sample consisted of 11 studies in which students were taught to use two or four cognitive strategies in the context of reciprocal teaching. All tests were in reading. Most of the researcher-developed tests consisted of reading paragraphs followed by multiple-choice questions, much like the standardized tests in reading. When summarization strategies were taught, however, the researcher-developed tests consisted of asking students to summarize a passage and the summaries were then coded. Other investigators used standardized tests to assess the results of instruction in summarization.

Across the three sets of studies, the median effect sizes were from .17 to .35 when standardized tests were used and from .70 to .95 when researcher-developed multiple-choice tests or summarization tests were used.

Summary. Although there is overlap in some of these samples, there are, at least, eight independent samples in Table 1. Across all these samples, the mean effect sizes were always higher when researcher-developed tests were used. These differences occurred across elementary, secondary and college-based studies. These differences occurred on standardized tests in reading, mathematics, social science, and science.

Results for studies that used both standardized tests and researcher-developed tests.

Another approach to studying this issue is to locate studies in which the researcher used both a standardized test and a researcher-developed test in the same study. Kulik et al (1990) located and reported on four such studies in their review of studies in Mastery Learning. Three studies were elementary education and one was in college. The median effect size was .06 when standardized tests were used and .25 when researcher-developed tests were used. In three of these studies the standardized tests were in mathematics; in the fourth study the standardized test was in reading, Rosenshine and Meister (1994) located nine such studies where both researcher-developed and standardized tests were used to assess reading comprehension. They found that the median effect size was .55 when standardized tests were used and .84 when researcher-developed tests were used,

Quality of the studies.

Some of the reviewers of these studies attended to the quality of the studies and only accepted studies that met their criteria. Other reviewers didn't discuss quality of the studies. But if we only accept the reviews that only accepted high-quality studies, the pattern is unchanged: the average effects sized on researcher-developed tests were higher than the effect sizes on each one of the remaining reviews.

Summary of results.

Eleven independent reviews of research were found of studies where attempts were made to improve student achievement. Some investigators used standardized-tests to measure student gains in achievement and some investigators used researcher-developed tests. Table 1 shows that in every one of the 11 reviews, the effect sizes were larger when researcher-developed tests were used to measure the effect of the intervention. These studies took place in seven different instructional formats (e.g. cooperative learning) and, in each format; effect sizes were higher when researcher-developer tests were used. These same results were also obtained in studies where cognitive strategies were taught in traditional settings. These same results favoring researcher-developed tests were obtained in all grade levels and in all subject areas that were tested. Finally, in studies in reading where both a researcher-developed and a standardized test were used the effect sizes were consistently larger when the researcher-developed test was used.

Discussion

Most of the authors of these eleven reviews did not discuss why these differences in results occurred. The major purpose of this review is to encourage further study on why greater gains were obtained when researcher-developed tests were used. A few points on this topic are discussed here.

1. Differences in content tested.

Gersten et al., (2009) have argued that the researcher-developed tests are closer to the mathematical skills that were taught in these studies, and the standardized-tests are more distant measures. In other words, the content assessed by the standardized test in mathematics or a science might be different from the content taught in the classroom. This topic would seem to merit further research. In reading comprehension, however, content is relatively less important and the focus is on the skill of comprehending new text. Still, the same pattern favoring researcher-developed tests was obtained.

2. Differences between standardized tests and researcher-developed tests in reading.

Alfassi (1998) specifically studied the differences between her researcher-developed test and a standardized test. Alfassi conducted an experimental study using reciprocal teaching. She obtained significant differences favoring the experimental group on the researcher-developed tests but not on the standardized tests. She then attempted to explore the causes of these differences.

Alfassi computed separate correlations between scores on her researcher-developed tests and scores on the Gates-MacGinitie Reading Tests, for experimental and control groups, at three points: pre-intervention, post-intervention and follow-up. All six

correlations ranged from .23 to .39. These relatively low correlations suggest that the two tests were not measuring common abilities.

Alfassi also studied the differences between the passages that were read in the two tests. Alfassi noted that all the passages in the researcher-developed tests consisted of expository material that followed a topic-sentence-and-supporting-detail format. In contrast, the expository passages in the Gates-MacGinitie Reading Tests (MacGinitie & MacGinitie, 1989) were shorter and did not follow this topic-sentence-and-supporting-detail format. Answering questions on standardized tests in reading, then, required a great deal more searching and inference from the reader, and this type of searching was not included in the cognitive strategies taught in the reciprocal teaching format. Alfassi thus argued that the tests and questions she developed had a stronger correspondence to the reciprocal teaching treatment that she students received. In addition, Alfassi noted that 29 of the 48 questions on the Gates-MacGinitie passages required vocabulary knowledge, and vocabulary development is not part of the strategies that are taught in reciprocal teaching.

Alfassi used these results to argue that the reciprocal teaching treatment trained students to answer questions about passages that had a topic sentence and supporting detail format, and questions that did not require additional vocabulary knowledge to answer.

Rosenshine et al. (1996) also compared the passages and questions in the two types of tests. They also noted that the paragraphs in the researcher-developed tests used by Palincsar and Brown (1982) were more "considerate" (Armbruster, 1984). The paragraphs that were developed by Palinscar tended to be organized in a main-idea-and-

supporting-detail format. In contrast, the paragraphs and passages in the standardized tests they inspected "did not have such a clear text structure" (Rosenshine et al., p.1996). Rosenshine et al (1996) also noted that students needed to have additional background knowledge in order to answer many of the questions on the standardized tests they inspected. On the standardized tests there were questions that asked why some words were italicized, questions that required additional vocabulary knowledge, questions that required inference beyond the text, questions that asked why a story had been written, and questions that asked where a passage might have appeared (p. 197). None of these types of questions appeared in the researcher-developed tests they inspected.

In summary, both authors noted that the expository passages in the standardized tests in they inspected did not have the main-idea-and supporting-detail format that appeared in the passages in researcher-developed tests. The questions in the standardized tests also required more inference and more vocabulary knowledge. Unfortunately, no articles were where the authors compared researcher-developed tests and standardized tests in science and mathematics. Such analyses would be useful.

Suggestions for additional research.

1. Study successful interventions more closely.

There may be interventions within each of these eleven meta-analyses that were particularly effective in raising student scores on a standardized test. It might be valuable to study these programs to see if we might identify instructional elements that distinguish those studies from studies that were less successful.

Both Alfassi (1998) and Rosenshine et al., (1996) noted that when students are learning to use cognitive strategies they almost always read considerate text, (Armbruster, 1984), that is, text with explicit topic sentences. But each of them noted that passages in the standardized tests they inspected were less considerate; the passages required more inference from the reader. Perhaps different instructional strategies are needed to raise student scores when reading inconsiderate text.

A study by Bereiter and Bird (1995) might be useful here. Bereiter and Bird (1985) first asked adult expert readers to think aloud while reading text that was difficult and often ambiguous. This procedure is a form of expert-novice research, where one attempts to identify the particular strategies that experts in a domain are using. Bereiter and Bird identified four fix-up strategies used by skilled readers when they encountered difficulty in comprehension of text. The four were:

1. Restatement of confusing text in simpler or more familiar terms
2. Backtracking and rereading
3. Attempting to identify the relationships between sections of the text
4. Formulating the difficulty as a problem

Students then received systematic instruction in using these four strategies. The instructors first identified, then modeled, and then explained these four strategies. Then they provided students with practice in identifying and applying the strategies. The instruction resulted in significantly increased their use of these strategies by the students and significant gains in reading comprehension on a standardized test.

Similarly, in Anderson's (1995) Adolescent Literacy Project (ALP) program, (1995). Low-scoring adolescent readers, studied difficult text, in a group, and discussed

strategies for understanding the text. Students in these programs have outscored control students on standardized tests.

The above two studies, in reading, provide an approach that is explicitly directed toward helping students read inconsiderate test, -- the type of text typically found on standardized tests in reading. In the first example, the processes of expert readers were studied. In the second example, struggling readers worked together to solve problems they faced when they read inconsiderate test. Perhaps either of these two approaches could be used in subject areas in science and social science.

2. Compute correlations between scores on researcher-developed tests and standardized-tests

The consistent differences in gains between researcher-developed tests and standardized-tests suggest that scores on these two types of tests have low correlations with each other. Alfassi (1998) computed correlations between student scores on her tests in reading and student scores on an achievement test in reading and found that these correlations ranged from .23 to .39. These low correlations suggest that the two tests are not measuring the same thing.

. It would be valuable if researchers were to go to completed studies, in all content areas, where both a researcher-developed test and a standardized-test were used, and, where possible, compute the correlations. It would also be valuable if a standardized test was used in future studies that used researcher-developed tests and correlations between scores on the two tests were computed.

It might also be valuable, where possible, to compute correlations between scores on state-wide tests and standardized-tests. Unfortunately, correlations cannot be

computed between student scores on tests administered by the National Assessment of Educational Progress (NAEP) and state-wide tests because individual students only take a section of a NAEP tests and therefore there are no individual scores on the NAEP tests in reading or mathematics, but correlations between state-wide tests and standardized-tests would be useful.

3. Compare items on standardized tests and researcher-developed tests.

Only two studies were found where the authors compared the format, passages, and questions on standardized tests with format, passages, and questions on researcher-developed tests, and both studies were in reading. There may be value in applying these analyses to other content areas such as mathematics and science and analyzing the differences in format and questions between researcher-developed tests and standardized-tests.

Students are presented with new passages followed by multiple-choice questions in both state-wide tests in reading comprehension and the NAEP tests in reading comprehension. Yet, Nichols, Glass, and Berliner (2006) have shown students in individual states have made much larger gains in reading and in mathematics on their state-wide tests than they did on the NAEP tests. It would be very useful to study if there were differences in how the reading passages were written? Were there differences in the format of the questions? Were there differences in the amount of background knowledge that was required to answer some questions? The phenomena of gains in state-wide tests and no gains on the NAEP tests have persisted over a number of years. Therefore, it would seem particularly useful to compare items, passages, and formats between the state-wide tests and the NAEP tests.

4. Present separate results for researcher-developed tests and standardized-tests in meta-analyses.

The meta-analyses assembled here have shown consistent differences in results between standardized tests and researcher-developed tests. These differences have occurred at all grade levels and in a variety of subject areas. These differences in effect size, and the correlations computed by Allessi (1990) suggest that the two types of tests may be measuring different things.

But in some of the meta-analyses, the reviewers published a mean effect size for standardized tests, a mean effect size for researcher-developed tests, and then they presented an overall mean effect size. But presenting an overall mean effect size may be misleading. For example, if one is studying cooperative learning and the median effect size was .20 when standardized tests were used and .60 when researcher-developed tests were used, it now seems inappropriate to present an overall, average effect size of .40 but reporting an average effect size in these cases can be misleading. It is possible that school personnel would adopt a program if they read, in a review of research, that the average effect size was .40, but would not adopt the same program if they knew that the average effect size for standardized tests was only .20.

Additional Comments.

Teaching to the test.

It is hard to argue that there was teaching to the test in the experimental groups when researcher-developed tests were used because the content was independent of the

intervention in these studies. For example, students in both the experimental and control groups studied the same material during Mastery Learning, the Personalized System of Instruction or Cooperative Learning, The only difference is that experimental students learned the material in a cooperative learning or a Mastery Learning setting. So it seems unlikely that there was teaching to the test in studies where researcher-developed tests were used.

In reciprocal teaching (Palinscar & Brown, 1984) students in the experimental groups practices learning strategies, and the researcher-developed tests required students to read and answer questions about new passages. All of the cognitive strategy interventions required students to read new passages and answer questions. So there was no test to be taught to.

Responsive to instruction?

A consistent finding in every one of the eleven meta-analyses in Table 1 is that every intervention was more effective when a researcher-developed test was used and was less effective when a standardized-test was used to measure achievement. In light of these consistent findings, do we need to change instruction in ways that might raise scores on standardized tests or do we need to change the standardized tests to make them more responsive to instruction?

Change the instruction. But even if it is argued that we should change instruction, we don't know what these changes might be. Consider cooperative learning (Slavin, 1990). It has been consistently shown that the effect sizes for cooperative learning were higher when researcher-developed test were used. Cooperative learning is a procedure

that involves students teaching others in their group, in settings where group scores are part of their grade. Both experimental and control groups always study the same material and take the same test. But there is nothing in cooperative learning that suggests how it might be modified to achieve higher scores on standardized tests.

The instructional procedures in reciprocal teaching (Palinscar & Brown, 1984) include students asking each other questions about a common passage, asking each other to clarify or to summarize or to predict what happen next. The effect sizes were much higher when researcher-developed tests were used. But there is nothing in reciprocal teaching that suggests how reciprocal teaching instruction and practice might be modified in order to obtain higher scores on standardized tests.

It becomes a very large problem if we argue that all of the interventions in Table 1 are inadequate and need to be modified in order to obtain larger gains on standardized-tests.

Change the standardized tests. On the other hand, across all these meta-analyses, the researcher-developed tests have been much more responsive to the instruction than the standardized-tests. A consistent finding in studies of cognitive strategy instruction has been that the typical cognitive strategy instruction, practiced with considerate text, has been more effective for raising student scores on researcher-developed tests than it is for standardized tests --tests which Alfassi (1998) says use inconsiderate text and questions that require additional vocabulary and inference beyond the test.

Perhaps the standardized tests --at least those in reading --should be changed to resemble the researcher-developed tests in reading. That is, more considerate passages and questions might be used in standardized tests.

The above discussion was limited to tests of reading comprehension. We do not how the researcher-developed tests and standardized-tests differed in science and mathematics. At any rate, we are left with the finding that effect sizes were greater when researcher-developed tests were used and we're unclear about the implications of this finding.

Conclusions.

1. The consistent gains when researcher-developed tests were in intervention studies show that the researcher-developed tests are responsive to instruction.

2. The small gains when standardized-tests were used shows that the standardized-tests were much less responsive to instruction.

3. These findings and the work of Allessi (1990) suggest that gains for the two types of tests have only low correlations. Therefore, one cannot expect that increased emphasis on the procedures used in these intervention studies will lead to strong gains on standardized-tests.

4. If standardized-tests are not very responsive to instruction, should they be considered the gold standard? Should our goal be that of trying to obtain higher scores on tests that are not very responsive to instruction?

References

Alfassi, M. (1998). Reading for meaning: the efficacy of reciprocal teaching in fostering reading comprehension in high school students in remedial reading classes. American Educational Research Journal, 35(2), 309-332.

Bredderman, T. (1982). What research says: Activity science-the evidence shows it matters. *Science and Children, 20*(1), 39-41. (ERIC Document Reproduction Service No. ED 216 870)

Gersten, R., Chard, D. J., Jayanthi, M.., Baker, S. K., Morphy, P. & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-Analysis of instructional components *Review of educational research*, 79, 1202-1242.

Kulik, C. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research,* 60, 265-299.

Kulik, J. A., Kulik, C. C., & Cohen, P. A. (1979). A meta-analysis of outcome studies of Keller's Personalized System of Instruction. *American Psychologist*, 34, 307-318.

Kulik, J.A., & Kulik, C.C. (1982). Effects of ability grouping on secondary school students: A meta-analysis of evaluation findings. *American Educational Research Journal,* 19, 415-428.

Lou, Y., Abrami, P., Spence, J., Poulsen, C., Chambers, B., & d'Apollonia, S, (1996). Within-class grouping: A meta-analysis. Review of Educational Research, 66, 423-458.

Nichols, S. L., Glass, G. V, & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives*, 14(1). Retrieved, October 1, 2009, from http://epaa.asu.edu/epaa/v14n1/.

Palincsar, A. S. & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and instruction, 2,* 117-175.

Rosenshine, B. & Meister, C. (1994). Reciprocal teaching: A review of the research. *Review of Educational Research*, 64, 479-531.

Rosenshine, B., Meister, C., & Chapman S., (1996). Teaching students to generate questions: A review of the intervention studies. *Review of educational research*, 66, 181-221.

Slavin, R.E. (1990). Cooperative learning: Theory, research, and practice. New Jersey: Prentice Hall.

Slavin, R. E. (1987). Mastery learning reconsidered. Review of educational research, 57, 175-213.

Swanson, H. L. & Hoskyn, M. (1998). Experimental intervention research on students with learning disabilities: A meta-Analysis of treatment outcomes, *Review of educational research* 68, 277-321.

Table 1. Mean effect-sizes for researcher-developed tests and standardized tests in 11 studies.

Study and subject area.	**Average effect sizes for researcher-developed tests (number of studies).**	**Average effect sizes for standardized tests (number of studies).**
Tests in a single subject		
Bredderman (1983). Activity-based elementary science.	.38 (30 studies)	.25 (22 studies).
Gersten et al. (2009). Mathematics for students with learning disabilities,	.73 (31 effect sizes)	.10 (10 effect sizes).
Rosenshine & Meister (1994). Reciprocal teaching-Reading	.88 (7)	.32 (11)
Rosenshine, Meister, & Chapman, (1996). Reading	.70 (12)	.35 (7)
Tests in different subjects		
Kulik . et al. (1982). Ability grouping in math, science, social science.	.11 (9)	.10 (42)
Lou et al. (1996) Small group learning in math, science, reading, or language arts.	.34 (21) researcher-developed tests. .42 (16) teacher-made tests.	.07 (61)
Kulik et al. (1979). Personalized system of instruction in science, math, or social science,	..52 (57)	.30 (9)
Kulik et al. (1990). Effectiveness of mastery learning programs: A	.65 (31) Mastery Learning	.

meta-analysis. Science and social science.		.10 (2)
Slavin (1987) Mastery learning in science, math, or reading.	.26 (8)	.04 (7)
Slavin (1990) Cooperative learning	.30 (12)	.20 (18)
Swanson & Hoskyn (1998) Students with learning disabilities	.86 (166 effect sizes)	.62 (90 effect sizes)