Abstract Title Page
Not included in page count.

Title:

Effective Educational Programs: Meta-Findings from the Best Evidence Encyclopedia

Author(s):

Robert E. Slavin
Johns Hopkins University
-and-
University of York

Cynthia Lake
Johns Hopkins University

Abstract Body
Limit 5 pages single spaced.

Background/context:

In recent years, there has been a great deal of interest in systematic reviews of research on educational programs and practices. In particular, the US What Works Clearinghouse, the UK EPPI-Centre, and the International Campbell Collaboration have produced series of reviews summarizing the effects of educational programs on student achievement. In each case, these reviews justify and establish procedures for searching the literature, standards for including studies in the review, methods for computing and then pooling effect sizes, and rules for characterizing findings for particular programs as strongly, moderately, or minimally indicating positive effects. Reviewers generally seek to end up with readily interpretable summaries, modeled on *Consumer Reports*, that inform educators about the confidence they might place in the likely impact of each program on student learning, with educator-friendly descriptions of the meaning of these findings. Although the reviews are scientific contributions, they are also clearly intended to inform practice and policy, and are a key element in the broader movement toward evidence-based reform in education.

One of the major series of reviews in elementary and secondary education is the Best Evidence Encyclopedia, or the BEE. This series of reviews, produced at Johns Hopkins University and the University of York (England), uses methods similar to those of the What Works Clearinghouse, but with more of a focus on large, lengthy studies with well-matched or randomized control groups and with measures not inherent to treatments (see details below, and Slavin, 2008). The BEE is systematically reviewing experimental research in many subjects in the preschool, elementary, and secondary grades, and has recently completed reviews of experimental studies of alternative textbooks, computer assisted instruction, instructional process/professional development programs, and combinations of these. The main reviews are as follows: *Effective programs in elementary mathematics: A best-evidence synthesis* (Slavin & Lake, 2008), *Effective programs in middle and high school mathematics* (Slavin, Lake, & Groff, in press), *Effective beginning reading programs: A best-evidence synthesis* (Slavin, Lake, Chambers, Cheung, & Davis), *Beyond the basics: Effective reading programs for the upper elementary grades* (Slavin, Lake, Cheung, & Davis, 2008), and *Effective reading programs for middle and high schools: A best evidence synthesis* (Slavin, Cheung, Groff, & Lake, 2008).

Collectively, these five reviews examined thousands of studies and found a total of more than 400 that met the inclusion standards (from 33 to 102 studies met the inclusion standards in the various reviews).

Up to now, findings for systematic reviews have largely been restricted to the reviews themselves, with few cases in which lessons learned across many reviews using similar methods can be synthesized. The completion of the Best Evidence Encyclopedia reading and math reviews permits a first opportunity to describe both substantive and methodological patterns across a broad set of studies involving all elementary and secondary grades, reviewed using a common set of review procedures.

Procedures Used in the Best-Evidence Encyclopedia. Although procedures have evolved somewhat over time, BEE reviews have used consistent core standards for literature searches, study inclusion, and computation and pooling of effect sizes. These are described in the following sections.

The review methods are adaptations of a technique called best evidence synthesis (Slavin, 1986). Best-evidence syntheses seek to apply consistent, well-justified standards to identify unbiased, meaningful information from experimental studies, discussing each study in some detail, and pooling effect sizes across studies in substantively justified categories. The method is very similar to meta-analysis (Cooper, 1998; Lipsey & Wilson, 2001), adding an emphasis on narrative description of each study's contribution and limiting the review to studies meeting the established criteria. It is also very similar to the methods used by the What Works Clearinghouse (2008), with a few important exceptions noted in the following sections. See Slavin (2008) for an extended discussion and rationale for the procedures used in all of these reviews.

Literature Search Procedures

A broad literature search was carried out in an attempt to locate every study that could possibly meet the inclusion requirements. Electronic searches were made of educational databases (JSTOR, ERIC, EBSCO, Psych INFO, Dissertation Abstracts) using different combinations of key words (for example, "elementary students", and "reading/mathematics achievement") and the years 1970-2008. Results were then narrowed by subject area (for example, "reading intervention," "mathematics program," "educational software," "academic achievement," "instructional strategies"). In addition to looking for studies by key terms and subject area, we conducted searches by program name. Web-based repositories and education publishers' websites were also examined. We attempted to contact producers and developers of reading programs to check whether they knew of studies that we had missed. Citations from other reviews in the same area were further investigated. We also conducted searches of recent tables of contents of key journals. Citations of studies appearing in the studies found in the first wave were also followed up. Studies meeting the selection criteria were included if they were published from 1970 to the present. Studies that met an initial screen for germaneness and basic methodological characteristics (e.g., they had a control group and a duration of at least 12 weeks) were then read by at least two of the present authors, always including the first and second author. Any disagreements in coding were resolved by discussion and by seeking advice from other authors.

Effect Sizes

In general, effect sizes were computed as the difference between experimental and control individual student posttests after adjustment for pretests and other covariates, divided by the unadjusted posttest control group standard deviation. If the control group SD was not available, a pooled SD was used. Procedures described by Lipsey & Wilson (2001) and Sedlmeier & Gigerenzor (1989) were used to estimate effect sizes when unadjusted standard deviations were not available, as when the only standard deviation presented was already adjusted for covariates or when only gain score SD's were available. If pretest and posttest means and SD's were presented but adjusted means were not, effect sizes for pretests were subtracted from effect sizes for posttests.

Effect sizes were pooled across studies for each program and for various categories of programs. This pooling used means weighted by the final sample sizes, computed as twice the smaller of the experimental or control number of students. The reason for using weighted means is to recognize the greater strength of large studies, as the previous reviews and many others

have found that small studies tend to overstate effect sizes (see Rothstein, Sutton, & Borenstein, 2005; Slavin, 2008; Slavin & Smith, 2008). A cap weight of 2500 students was used to avoid having very large studies dominate the pooled means.

Criteria for Inclusion

Criteria for inclusion of studies in this review were as follows.

1. The studies evaluated programs in the relevant subject and grade levels. Studies of variables, such as use of ability grouping, block scheduling, or single-sex classrooms, were not reviewed.
2. The studies evaluated programs intended for all children. Remedial, preventive, and special education programs are being reviewed in a separate synthesis (Slavin et al., forthcoming).
3. The studies compared children taught in classes using a given program with those in control classes using an alternative program or standard methods.
4. Studies could have taken place in any country, but the report had to be available in English.
5. Random assignment or matching with appropriate adjustments for any pretest differences (e.g., analyses of covariance) had to be used. Studies without control groups, such as pre-post comparisons and comparisons to "expected" scores, were excluded.
6. Pretest data had to be provided, unless studies used random assignment of at least 30 units (individuals, classes, or schools) and there were no indications of initial inequality. Studies with pretest differences of more than 50% of a standard deviation were excluded because, even with analyses of covariance, large pretest differences cannot be adequately controlled for as underlying distributions may be fundamentally different (Shadish, Cook, & Campbell, 2002).
7. The dependent measures included quantitative measures of academic performance, such as standardized reading and math measures. Experimenter-made measures were accepted if they were comprehensive measures of reading or math, which would be fair to the control groups, but measures of objectives inherent to the program (but unlikely to be emphasized in control groups) were excluded. Studies using measures inherent to treatments, such as those made by the experimenter or program developer, or measures of skills taught only in the treatment group, have been found to be associated with much larger effect sizes than are measures that are independent of treatments (Slavin & Madden, 2008), and for this reason, effect sizes from treatment-inherent measures were excluded. The exclusion of measures inherent to the experimental treatment is a key difference between the procedures used in the present review and those used by the What Works Clearinghouse.
8. A minimum study duration of 12 weeks was required. This requirement was introduced to focus the review on practical programs intended for use for the whole year, rather than brief investigations. Brief studies may not allow programs to show their full effect. On the other hand, brief studies often advantage experimental groups that focus on a particular set of objectives during a limited time period while control groups spread that topic over a longer period. Studies with brief treatment durations that measured outcomes over periods of more than 12 weeks after implementation began were

included, however, on the basis that if a brief treatment has lasting effects, it should be of interest to educators. The 12-week criterion has been consistently used in all of the systematic reviews done previously by the current authors (i.e., Cheung & Slavin, 2005; Slavin & Lake, 2008; Slavin et al., 2008).

9. Studies had to have at least two teachers and 15 students in each treatment group.

Purpose/objective/research question/focus of study:

The purpose of the proposed paper is to synthesize both substantive and methodological findings across the five main Best Evidence Encyclopedia reviews of reading and math programs in grades K-12. The paper will consider the following research questions:

1. Across subjects and grade levels, what effect sizes are associated with variations in a) textbooks, b) computer-assisted instruction, c) instructional process programs, and d) combinations of these? What subcategories within these types of interventions are associated with positive effects?
2. How do summary outcomes of various types of programs vary across reading and math, and across elementary and secondary grades?
3. Across subjects and grade levels, how do effect sizes differ according to the following methodological criteria:
   a. Use of random assignment
   b. Sample size
   c. Duration
   d. Use of standardized measures

Setting:
n/a

Population/Participants/Subjects:
n/a

Intervention/Program/Practice:
n/a

Research Design:
Synthesis of systematic reviews

Data Collection and Analysis:
The paper will present sample size-weighted effect sizes from each of the BEE reviews, broken down by type of program, by use of random versus matched assignment, by study duration, and by type of outcome measure. It will present unweighted mean effect sizes to investigate effects of sample size. Effect sizes will be pooled across categories to examine patterns by grade level and by subject, and to investigate interesting contrasts suggested by the main comparisons.

Findings/Results:
Preliminary findings across the five reviews support the following conclusions:

1. Across all subjects and grade levels, instructional process approaches are associated with the most positive effect sizes. Within this category, cooperative learning programs such

as *PALS* (reading and math), *Classwide Peer Tutoring* (reading and math), *Student Teams Achievement Divisions* (math), *IMPACT* (math), and *Cooperative Integrated Reading and Composition* (reading), are consistently associated with the largest effect sizes (and the largest numbers of studies with positive effect sizes). Other instructional process programs with notably positive effect sizes include programs that teach metacognitive learning strategies and those that introduce effective classroom management and motivation strategies. Programs that combine instructional processes with curriculum, especially *Success for All* (reading), *Direct Instruction* (reading and math), and *Team Assisted Instruction* (math) obtain particularly large and frequently replicated positive outcomes.

2. Effects of traditional computer-assisted instruction (CAI) (e.g., *Jostens/Compass Learning*, *CCC/Successmaker*, and similar programs) are modest in math and near zero in reading. However, programs that combine CAI with instructional process approaches, such as *Read 180*, and programs that use technology to improve teachers' classroom instruction, such as *Reading Reels*, have been associated with positive reading outcomes.
3. Studies comparing alternative core and supplemental textbooks (e.g., Scott Foresman, Houghton Mifflin, Everyday Math) find near-zero effect sizes at all grade levels and in both subjects. On measures not inherent to the curricula themselves, reform-oriented textbooks (such as math texts supported by the National Science Foundation), traditional texts, and back-to-basics texts (e.g., Saxon Math) have rarely been found to differ from control groups in student outcomes.
4. Studies that used random assignment to conditions and those that use matched comparisons report nearly identical effect sizes.
5. Studies with small sample sizes tend to report larger effect sizes than those with large sample sizes. For this reason, current BEE reviews weight outcomes by sample size.
6. Studies with brief durations (at least 12 weeks but less than a year) report somewhat larger effect sizes than longer studies (more than a year).

Conclusions:

The "meta-findings" across the five Best Evidence Synthesis reviews suggest that strategies likely to improve student learning are those that improve the quality of daily instruction, increase students' active participation in the classroom, and help students learn metacognitive skills. Consistently successful programs, such as cooperative learning, teaching of metacognitive skills, and improved management and motivation approaches, as well as comprehensive programs such as *Success for All* and *Direct Instruction*, all emphasize extensive professional development, typically including multi-day workshops, in-class followup, and clear guidance and extensive supportive materials for teachers. Technology can be effective to the degree that it also supports active instruction, cooperative learning, and improving classroom instruction. Changing curriculum or textbooks is rarely an effective strategy in itself, but may be an important element of comprehensive approaches that also incorporate instructional processes.

Methodological patterns were also consistent across subjects and grade levels. Surprisingly, random assignment never made an important difference in effect sizes. Far more important were sample size, duration, and use of measures not inherent to treatments.

Appendixes

Not included in page count.

Appendix A. References

References are to be in APA format. (See APA style examples at the end of the document.)

Chambers, E. A. (2003). *Efficacy of educational technology in elementary and secondary classrooms: A meta-analysis of the research literature from 1992-2002*. Unpublished doctoral dissertation, Southern Illinois University at Carbondale.

Cheung, A., & Slavin, R.E. (2005). Effective reading programs for English language learners and other language minority students. *Bilingual Research Journal, 29* (2), 241-267.

Cooper, H. (1998). *Synthesizing research (3rd ed.).* Thousand Oaks, CA: Sage.

Deshler, D., Palincsar, A., Biancarosa, G., & Nair, M. (2007). *Informed choices for struggling adolescent readers.* Newark, DE: International Reading Association.

Kulik, J. A. (2003). *Effects of using instructional technology in elementary and secondary schools: What controlled evaluation studies say. SRI Project Number P10446.001*. Arlington, VA: SRI International.

Lipsey, M.W., & Wilson, D.B. (2001). *Practical meta-analysis.* Thousand Oakes, CA: Sage

Murphy, R., Penuel, W., Means, B., Korbak, C., Whaley, A., & Allen, J. (2002). *E-DESK: A review of recent evidence on discrete educational software*. Menlo Park, CA: SRI International.

Rothstein, H.R., Sutton, A.J., & Borenstein, M. (Eds.) (2005). *Publication bias in meta-analysis: Prevention, assessment, and adjustments.* Chichester, UK: John Wiley.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309-316.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Slavin, R. E. Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher*, 1986, 15, (9), 5-11.

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

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.

Slavin, R., & Lake, C. (2008). Effective programs in elementary mathematics; A best-evidence synthesis. *Review of Educational Research, 78* (3), 427-515.

Slavin, R.E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2008). *Effective beginning reading programs: A best-evidence synthesis*. Baltimore, MD: Center for Research and Reform in Education, Johns Hopkins University.

Slavin, R.E., Lake, C., Cheung, A., & Davis, S. (2008). *Beyond the basics: Effective reading programs for the upper elementary grades*. Baltimore, MD: Center for Research and Reform in Education, Johns Hopkins University.

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

Slavin, R.E., & Madden, N. A. (2008, March). *Understanding bias due to measures inherent to treatments in systematic reviews in education.* Paper presented at the annual meetings of the Society for Research on Educational Effectiveness, Crystal City, Virginia.

Slavin, R.E., & Smith, D. (2008, March). *Effects of sample size on effect size in systematic reviews in education*. Paper presented at the annual meetings of the Society for Research on Educational Effectiveness, Crystal City, Virginia.

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