Robert E. Slavin

Director, Center for Data-Driven Reform in Education
John Hopkins University

Chairman, Board of Directors
Success for All Foundation

Nominated by
Charles Read

Robert E. Slavin
Director, Center for Data-Driven Reform in Education, John Hopkins University
Chairman, Board of Directors, Success for All Foundation

Robert Slavin is currently Director of the Center for Data-Driven Reform in Education at Johns Hopkins University and Chairman of the Success for All Foundation. He received his B. A. in Psychology from Reed College in 1972, and his Ph.D. in Social Relations in 1975 from Johns Hopkins University. Dr. Slavin has authored or co-authored more than 200 articles and 20 books, including *Educational Psychology: Theory into Practice* (Allyn & Bacon, 1986, 1988, 1991, 1994, 1997, 2000, 2003), *Cooperative Learning: Theory, Research, and Practice* (Allyn & Bacon, 1990, 1995), *Show Me the Evidence: Proven and Promising Programs for America's Schools* (Corwin, 1998), *Effective Programs for Latino Students* (Erlbaum, 2000), and *One Million Children: Success for All* (Corwin, 2001). He received the American Educational Research Association's Raymond B. Cattell Early Career Award for Programmatic Research in 1986, the Palmer O. Johnson award for the best article in an AERA journal in 1988, the Charles A. Dana award in 1994, the James Bryant Conant Award from the Education Commission of the States in 1998, the Outstanding Leadership in Education Award from the Horace Mann League in 1999, and the Distinguished Services Award from the Council of Chief State School Officers in 2000.

Submitted by: Charles Read

VITA

Robert Edward Slavin
Center for Data-Driven Reform in Education
Johns Hopkins University
200 W. Towsontown Blvd.
Baltimore, MD 21204
410-616-2310
e-mail: rslavin@cddre.org

Education

B.A.	Reed College, Portland, Oregon, 1972 Major: Psychology Minor: Education
Ph.D.	Department of Social Relations, Johns Hopkins University, Baltimore, Maryland, 1975

Experience

2004 - present	Director, Center for Data-Driven Reform in Education, Johns Hopkins University, Baltimore, MD
1997 - present	Chairman, Board of Directors, Success for All Foundation, Baltimore, MD
1994 - 2004	Principal Research Scientist and Co-Director, Center for Research on the Education of Students Placed at Risk, Johns Hopkins University
1989 - 1994	Principal Research Scientist and Co-Director, Early and Elementary Program, Center for Research on Effective Schooling for Disadvantaged Students, Johns Hopkins University
1985 - 1990	Director, Elementary School Program, Center for Research on Elementary and Middle Schools, Johns Hopkins University
1978 - 1985	Research Scientist, Center for Social Organization of Schools, Johns Hopkins University
1975 - 1978	Associate Research Scientist, Center for Social Organization of Schools, Johns Hopkins University.
1972 - 1973	Teacher, Aloha Children' s Center (children with mental disabilities), Aloha, OR.
1970 - 1971	Student teacher (Social Studies, Psychology, Minority Studies, U.S. History), Aloha High School, Aloha, OR

Awards and Scholarships

Commendation for Excellence in Scholarship, Reed College, 1972.

NIMH Training Grant, Johns Hopkins University, 1973-1975.

Raymond B. Cattell Early Career Award for Programmatic Research, American Educational Research Association, 1985.

Distinguished Scholar, AERA Committee on the Role and Status of Minorities in Educational Research, 1987.

Palmer 0. Johnson Award (best article in an AERA journal in 1986), American Educational Research Association, 1988.

Distinguished Achievement Award for Excellence in Educational Journalism, Educational Press Association of America, 1988.

Award for Distinguished Scientific Contributions to Psychology, Maryland Psychological Association, 1988.

Parents Magazine "As They Grow" award in education, 1994.

Charles A. Dana Award, 1994.

James Bryant Conant Award, Education Commission of the States, 1998.

Outstanding Educator Award, Horace Mann League, 1999.

Honorary Doctorate, University of Liège, Belgium, 1999.

Distinguished Service Award, Council of Chief State School Officers, 2000.

Distinguished Achievement Award for Excellence in Educational Publishing, *Educational Leadership*, 2003.

Professional Associations

American Educational Research Association
- Chair, Publications Committee, 1990-92
- Member-at-Large, AERA Council, 1996-99

Association for Supervision and Curriculum Development

International Association for the Study of Cooperation in Education
- Vice President, 1985-1988; President, 1988-1990

North American Editor, *Educational Research and Evaluation,* 1996-99.

Books

Slavin, R. E. *Using Student Team Learning.* Baltimore, MD: Johns Hopkins Team Learning Project, 1978. Second edition, 1980. Third edition, 1986. Fourth Edition, 1994

DeVries, D. L., Slavin, R. E., Fennessey, G. M., Edwards, K. J., & Lombardo, M. *Teams-Games-Tournament: The team learning approach.* Englewood Cliffs, NJ: Educational Technology Publications, 1980.

Slavin, R. E. *Cooperative learning in student teams: What research says to the teacher.* Washington, D.C.: National Education Association, 1982. Revised edition, 1987.

Slavin, R. E. *Student Team Learning: An overview and practical guide.* Washington, D.C.: National Education Association, 1983. Second edition, 1988.

Slavin, R. E. *Cooperative learning.* New York: Longman, 1983.

Slavin, R. E. *Research methods in education: A practical guide.* Englewood Cliffs, NJ: Prentice-Hall, 1984. Second edition, 1992.

Slavin, R. E., Sharan, S., Kagan, S., Hertz-Lazarowitz, R., Webb, C., & Schmuck, R. (Eds.), *Learning to cooperate, cooperating to learn.* New York: Plenum, 1985.

Slavin, R. E., Leavey, M. B., & Madden, N. A. *Team Accelerated Instruction - Mathematics.* Watertown, MA: Charlesbridge, 1986.

Slavin, R. E. *Educational psychology: Theory into practice.* Boston: Allyn & Bacon, 1986. Second edition, 1988. Third edition, 1991. Fourth Edition, 1994. Fifth Edition, 1997. Sixth Edition, 2000. Seventh edition, 2003.

Slavin, R. E. (Ed.) *School and classroom organization.* Hillsdale, NJ: Erlbaum, 1989.

Slavin, R. E., Karweit, N. L., & Madden, N. A. (Eds.) *Effective programs for students at risk.* Boston, MA: Allyn and Bacon, 1989.

Slavin, R. E. *Cooperative learning: Theory, research, and practice.* Boston: Allyn & Bacon, 1990. Second Edition, 1995.

Slavin, R. E., Madden, N. A., Karweit, N. L., Dolan, L. J., & Wasik, B. A. *Success for All: A relentless approach to prevention and early intervention in elementary schools.* Arlington, VA: Educational Research Service, 1992.

Slavin, R. E., Karweit, N. L., Wasik, B. A., & Madden, N. A. (Eds.) *Preventing early school failure: Research, policy, and practice.* Boston: Allyn & Bacon, 1994.

Slavin, R. E., Madden, N. A., Dolan, L. J., & Wasik, B. A. *Every child, every school: Success for All.* Thousand Oaks, CA: Corwin, 1996.

Slavin, R.E. *Education for all.* Lisse, The Netherlands: Swets & Zeitlinger, 1997.

Slavin, R.E., & Fashola, O.S. *Show me the evidence: Proven and promising programs for America's schools.* Thousand Oaks, CA: Corwin, 1998.

Slavin, R.E., & Calderón, M. (Eds.) (2001), *Effective Programs for Latino Students.* Mahwah, NJ: Erlbaum.

Slavin, R.E., & Madden, N.A. (Eds.) (2001). *One million children: Success for All.* Thousand Oaks, CA: Corwin.

Slavin, R.E., & Madden, N.A. (Eds.) (2001). *Success for All: Research and reform in elementary education.* Mahwah, NJ: Erlbaum.

Borman, G., Stringfield, S., & Slavin, R.E. (Eds.) (2001). *Title I: Compensatory education at the crossroads.* Mahwah, NJ: Erlbaum.

Morris, D., & Slavin, R. (2003). *Every child reading.* Boston: Allyn & Bacon.

Slavin, R.E. (2003). *Educational psychology: Theory into practice.* (7^{th} ed.). Boston: Allyn & Bacon.

Slavin, R.E. (2006). *Educational psychology: Theory into practice.* (8^{th} ed.). Boston: Allyn & Bacon.

Journal Articles

Slavin, R. E. A student team approach to teaching adolescents with special emotional and behavioral needs. *Psychology in the Schools*, 1977, 14, 77-84.

Slavin, R. E. Classroom reward structure: An analytic and practical review. *Review of Educational Research*, 1977, 47, 633-650.

Slavin, R. E. Using student learning teams to integrate the desegregated classroom. *Integrated Education*, 1977, 15, 56-58.

Slavin, R. E. Basic vs. applied research: A response. *Educational Researcher*, 1978, 7, 15-17.

DeVries, D. L., Edwards, K. J., & Slavin, R. E. Biracial learning teams and race relations in the classroom: Four field experiments using Teams-Games-Tournament. *Journal of Educational Psychology*, 1978, 70, 356-362.

Slavin, R. E. Student teams and comparison among equals: Effects on academic performance and student attitudes. *Journal of Educational Psychology*, 1978, 70, 532-538.

Slavin, R. E. Student teams and achievement divisions. *Journal of Research and Development in Education*, 1978, 12, 39-49.

DeVries, D. L., & Slavin, R. E. Teams-Games-Tournament: Review of ten classroom experiments. *Journal of Research and Development in Education*, 1978, 12, 28-38.

Slavin, R. E. Separating incentives, feedback, and evaluation: Toward a more effective classroom system. *Educational Psychologist*, 1978, 13, 97-100.

Slavin, R. E. Effects of biracial learning teams on cross-racial friendships. *Journal of Educational Psychology*, 1979, 71, 381-387.

Slavin, R. E., & Tanner, A. M. Effects of cooperative reward structures and individual accountability on productivity and learning. *Journal of Educational Research*, 1979, 72, 294-298.

Slavin, R. E. Integrating the desegregated classroom: Action speaks louder than words. *Educational Leadership*, 1979, 36 (5), 322-324.

Slavin, R. E., & Madden, N. A. School practices that improve race relations. *American Educational Research Journal*, 1979, 16, 169-180.

Slavin, R. E. Competence and competency-based graduation requirements. *Baltimore Sun*, January 11, 1979.

Slavin R. E. Cooperative learning. *Review of Educational Research*, 1980, 50, 315-342. Named a "citation classic" (cited at least 100 times) by Current Contents, 1988.

Slavin, R. E. Effects of student teams and peer tutoring on academic achievement and time on-task. *Journal of Experimental Education*, 1980, 48, 252-257.

Slavin, R. E. Effects of individual learning expectations on student achievement. *Journal of Educational Psychology*, 1980, 72, 520-524.

Slavin, R. E. Cooperative learning in teams: State of the art. *Educational Psychologist*, 1980, 15, 93-111.

Slavin, R. E. Cooperative learning and desegregation. *Journal of Educational Equity and Leadership*, 1981, 1, 145-161.

Slavin, R. E., Wodarski, J. L., & Blackburn, G. L. A group contingency for electricity consumption in master-metered apartments. *Journal of Applied Behavioral Analysis*, 1981, 14, 357-363.

Slavin, R. E., & Karweit, N. Cognitive and effective outcomes of an intensive student team learning experience. *Journal of Experimental Education*, 1981, 50, 29-35.

Hansell, S., Tackaberry, S. N., & Slavin, R. E. Cooperation, competition, and the structure of student peer groups. *Representative Research in Social Psychology*, 1981, 12, 46-61.

Slavin, R. E., & Oickle, E. Effects of cooperative learning teams on student achievement and race relations: Treatment x race interactions. *Sociology of Education*, 1981, 54, 174-180.

Beady, C. L., Slavin, R. E., & Fennessey, G. M. Alternative student evaluation structures and a focused schedule of instruction in an inner-city junior high school. *Journal of Educational Psychology*, 1981, 73, 518-523.

Hansell, S., & Slavin, R. E. Cooperative learning and interracial friendships. *Sociology of Education*, 1981, 54, 98-106.

Slavin, R. E. A case study of psychological research affecting classroom practice: Student Team Learning. *Elementary School Journal*, 1981, *82*, 5-17.

Slavin, R. E. A policy choice: Cooperative or competitive learning. *Character*, 1981, 2 (3), 1-6.

Karweit, N., & Slavin, R. E . Measurement and modeling choices in studies of time and learning. *American Educational Research Journal*, 1981, 18, 157-171.

Slavin, R. E . Cooperative learning: A research synthesis. *Educational Leadership*, 1981, 38, 655-660.

Beady, C. L., & Slavin, R. E. Making success available to all students in desegregated schools. *Integrated Education*, 1981, 18 (6), 28-31.

Karweit, N., & Slavin, R. E. Time-on-task: Issues of timing, sampling and definition. *Journal of Educational Psychology*, 1982, 74, 844-851.

Madden, N. A., & Slavin, R. E. Effects of cooperative learning on the social acceptance of mainstreamed academically handicapped students. *Journal of Special Education*, 1983, 17, 171-1 82.

Hollifield, J. H., & Slavin, R. E. Disseminating Student Team Learning through Federally funded programs: Appropriate technology, appropriate channels. *Knowledge*, 1983, 4, 576-589.

Slavin, R. E. Educational psychology for teachers. *Contemporary Psychology*, 1983, 28, 401-402.

Slavin, R. E. When does cooperative learning increase student achievement? *Psychological Bulletin*, 1983, 94, 429-445.

Slavin, R. E. A nation at risk: Realities and remedies. *Elementary School Journal*, 1983, 84, 131-138.

Madden, N. A., & Slavin, R. E. Mainstreaming students with mild academic handicaps: Academic and social outcomes. *Review of Educational Research*, 1983, 53, 519-569.

Slavin, R. E. Component building: A strategy for research-based instructional improvement. *Elementary School Journal*, 1984, 84, 255-269 .

Slavin, R. E., Madden, N., & Leavey, M. Effects of cooperative learning and individualized instruction on mainstreamed students. *Exceptional Children*, 1984, 50, 434-443.

Slavin, R. E., Leavey, M., & Madden, N. A. Combining cooperative learning and individualized instruction: Effects on student mathematics achievement, attitudes, and behaviors. *Elementary School Journal*, 1984, 84, 409-422.

Slavin, R. E. Students motivating students to excel: Cooperative incentives, cooperative tasks, and student achievement. Elementary School Journal, 1984, 85, 53-64. Reprinted in H. F. Clarizio, R. C. Craig, and W. A. Mehrens (Eds.), *Contemporary issues in educational psychology*. New York: Random House, 1986.

Slavin, R. E., Madden, N. A., & Leavey, M. Effects of Team Assisted Individualization on the mathematics achievement of academically handicapped and non-handicapped students. *Journal of Educational Psychology*, 1984, 76, 813-819.

Slavin, R. E. Meta-analysis in education: How has it been used? *Educational Researcher*, 1984, 13 (8), 6-15.

Slavin, R. E. A rejoinder to Carlberg et al. *Educational Researcher*, 1984, 13 (8), 24-27.

Slavin, R. E. Team Assisted Individualization: Cooperative learning and individualized instruction in the mainstreamed classroom. *Remedial and Special Education*, 1984, 5 (6), 33-42.

Slavin, R. E., & Karweit, N. Mastery learning and student teams: A factorial experiment in urban general mathematics classes. *American Educational Research Journal*, 1984, 21, 725-736.

Slavin, R. E., & Karweit, N. Effects of whole class, ability grouped, and individualized instruction on mathematics achievement. *American Educational Research Journal*, 1985, 22 (3), 351-367.

Slavin, R. E. Cooperative learning: Applying contact theory in desegregated schools. *Journal of Social Issues*, 1985, 41, 45-62.

Slavin, R. E. Learning together. *American Educator*, 1986, 10, (2), 6-13.

Slavin, R. E. Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher*, 1986, 15, (9), 5-11. Winner of the Palmer 0. Johnson Award (best article in an AERA journal in 1986), American Educational Research Association, 1988. Reprinted in W. R. Shadish and C. S. Reichardt (Eds.), *Evaluation Studies Review Annual* (Vol. 12). Newbury Park, CA: Sage, 1987.

Slavin, R. E. The Napa evaluation of Madeline Hunter's ITIP: Lessons learned. *Elementary School Journal*, 1986, 87, 165-171.

Slavin, R. E., & Madden, N. A. The integration of students with mild academic handicaps in regular classrooms. *Prospects* (Unesco), 1986, 16, 443-461.

Slavin, R. E. A theory of school and classroom organization. *Educational Psychologist*, 1987, 22, 89-108.

Slavin, R. E. Grouping for instruction in the elementary school. *Educational Psychologist*, 1987, 22, 109-127.

Slavin, R. E. Cooperative learning: Can students help students learn? *Instructor*, 1987 (March), 74-78.

Slavin, R. E. Developmental and motivational perspectives on cooperative learning: A reconciliation. *Child Development*, 1987, 58, 1161-1167. Article also published in Italian translation as Apprendimento cooperativo, una proposta di conciliazione fra prospettive evolutive e motivazionali. *Eta ' Evolutiva*, 1986, 24, 54-61.

Slavin, R. E. The Hunterization of America's schools. *Instructor*, 1987 (April), 56-58.

Slavin, R. E. Best-evidence synthesis: Why less is more. *Educational Researcher*, 1987, 4, 15-16.

Slavin, R. E. Ability grouping and its alternatives: Must we track? *American Educator*, 1987, 11 (2), 32-36, 47-48.

Slavin, R. E. Mastery learning reconsidered. *Review of Educational Research*, 1987, 57, 175-213.

Slavin, R. E. Taking the mystery out of mastery: A response to Guskey, Anderson, and Burns. *Review of Educational Research*, 1987, 57, 231-235.

Slavin, R. E. Cooperative learning in geometry: 180 degrees off. *Contemporary Psychology*, 1987, 32, 62-63.

Slavin, R. E. Cooperative learning: Where behavioral and humanistic approaches to classroom motivation meet. *Elementary School Journal*, 1987, 88, 29-37.

Slavin, R. E. Making Chapter 1 make a difference. *Phi Delta Kappan*, 1987, 69, 110-119. Winner of the Distinguished Achievement Award for Excellence in Educational Journalism, Educational Press Association of America, 1988.

Slavin, R. E. Grouping for instruction: Equity and effectiveness. *Equity and Excellence*, 1987, 23, 31-36.

Stevens, R. J., Madden, N. A., Slavin, R. E., & Farnish, A. M. Cooperative Integrated Reading and Composition: Two field experiments. *Reading Research Quarterly*, 1987, 22, 433-454.

Slavin, R. E. Combining cooperative learning and individualized instruction. *Arithmetic Teacher*, 1987, 35 (3), 14-16.

Slavin, R. E. Cooperative learning and the cooperative school. *Educational Leadership*, 1987, 45 (3), 7-13.

Slavin, R. E. Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research*, 1987, 57, 293-336.

Slavin, R. E. Ability grouping in elementary schools: Do we really know nothing until we know everything? *Review of Educational Research,* 1987, 57, 347-350.

Slavin, R. E. Cooperative revolution catches fire. *The School Administrator*, 1988 (January), 45, 9-13.

Slavin, R. E., Stevens, R. J., & Madden, N. A. Accommodating student diversity in reading and writing instruction: A cooperative learning approach. *Remedial and Special Education*, 1988, 9, 60-66.

Slavin, R. E. Winning isn't everything, but is it anything? *Contemporary Psychology*, 1988, 33, 151.

Slavin, R. E. Synthesis of research on grouping in elementary and secondary schools. *Educational Leadership,* 1988, 46 (1), 67-77.

Slavin, R. E. Cooperative learning and student achievement. *Educational Leadership*, 1988, 46 (2), 31-33.

Slavin, R. E. Back up the bucks with effective programs for Chapter 1 children. *R & D Preview*, 1988, 3 (2), 12.

Slavin, R. E. Class size and student achievement: Small effects of small classes. *Educational Psychologist*, 1989, 24, 99-110.

Slavin, R. E., & Madden, N. A. What works for students at risk: A research synthesis. *Educational Leadership*, 1989, 46 (5), 4-13. Winner of the 1989 prize for scholarly feature articles awarded by the Society of National Association Publications.

Madden, N. A., Slavin, R. E., Karweit, N. L., & Livermon, B. J. Success for All: Restructuring the urban elementary school. *Educational Leadership*, 1989, 46 (5), 14-18.

Slavin, R. E. On mastery learning and mastery teaching. *Educational Leadership*, 1989, 46 (7), 77-79.

Slavin, R. E. Comprehensive cooperative learning models for heterogeneous classrooms. *The Pointer*, 1989, 33 (2), 12-19.

Slavin, R. E. PET and the pendulum: Faddism in education and how to stop it. *Phi Delta Kappan*, 1989, 70, 752-758.

Slavin, R. E., Madden, N. A., Karweit, N. L., Livermon, B. J., & Dolan, L. Can every child learn? An evaluation of "Success for All" in an urban school. *Journal of Negro Education*, 1989, 58, 357-366.

Slavin, R. E. Research on cooperative learning: An international perspective. *Scandanavian Journal of Educational Research*, 1989, 33,231-243.

Slavin, R. E., Madden, N. A., & Stevens, R. J. Cooperative learning models for the 3 R's. *Educational Leadership*, 1989/1990, 47 (4), 22-28.

Slavin, R. E. Cooperative learning: Here to stay or gone tomorrow? *Educational Leadership*, 1989/1990, 47 (4), 22-28.

Slavin, R. E. Research on cooperative learning: Consensus and controversy. *Educational Leadership*, 1989/1990, 47 (4), 52-55.

Slavin, R. E. On making a difference. *Educational Researcher*, 1990, 19 (3), 30-34.

Slavin, R. E. General education under the Regular Education Initiative: How must it change? *Remedial and Special Education*, 1990, 11 (3), 40-61.

Slavin, R. E., Madden, N. A., Karweit, N. L., Livermon, B. J., & Dolan, L. Success for All: First year outcomes of a comprehensive plan for reforming urban education. *American Educational Research Journal*, 1990, 27, 255-278.

Slavin, R. E. On Success for All: Defining "success", defining "all". *Remedial and Special Education*, 1990, 11, (4), 60-61.

Slavin, R. E. Learning together. *American School Board Journal*, 1990, 177 (8), 22-23.

Slavin, R. E. Mastery learning re-reconsidered. *Review of Educational Research*, 1990, 60, 300-302.

Slavin, R. E. Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research*, 1990, 60, 471-499.

Slavin, R. E. Ability grouping in secondary schools: A response to Hallinan. *Review of Educational Research*, 1990, 60, 505-507.

Slavin, R. E. Ability grouping, cooperative learning, and the gifted. *Journal for the Education of the Gifted*, 1990, 14, 3-8.

Slavin, R. E. Cooperative learning and the gifted: Who benefits? *Journal for the Education of the Gifted*, 1990, 14, 28-30.

Slavin, R. E. A synthesis of research on cooperative learning. *Educational Leadership*, 1991, 48 (5), 71-82.

Stevens, R. J., Slavin, R. E., & Farnish, A. M. The effects of cooperative learning and direct instruction in reading comprehension strategies on main idea identification. *Journal of Educational Psychology*, 1991, 83, 8-16.

Slavin, R. E. IBM's "Writing to Read" program: Is it right for reading? *Phi Delta Kappan*, 1990, 72, 214-216.

Stevens, R. J., & Slavin, R. E. When cooperative learning improves the achievement of students with mild disabilities: A response to Tateyama-Sniezek. *Exceptional Children*, 1990, 57, 276-280.

Slavin, R. E. Class size and student achievement: Is smaller better? *Contemporary Education*, 1990, 42, 6-12.

Slavin, R. E. Group rewards make groupwork work. *Educational Leadership*, 1991, 48 (5), 89-91.

Slavin, R. E. Reading effects of IBM's "Writing to Read" program: A review of evaluations. *Educational Evaluation and Policy Analysis*, 1991, 13, 1-12.

Slavin, R. E. Are cooperative learning and untracking harmful to the gifted? *Educational Leadership*, 1991, 48 (6), 68-71.

Slavin, R. E. Chapter 1: A vision for the next quarter-century. *Phi Delta Kappan*, 1991, 72 (8), 586-592.

Madden, N. A., Slavin, R. E., Karweit, N. L., Dolan, L., & Wasik, B. Success for *all*. *Phi Delta Kappan*, 1991, 72 (8), 593-599.

Slavin, R. E., Madden, N. A., Karweit, N. L., Dolan, L., Wasik, B. A., Shaw, A., Mainzer, K. L., & Haxby, B. Neverstreaming: Prevention and early intervention as an alternative to special education. *Journal of Learning Disabilities*, 1991, 24, 373-378.

Slavin, R. E., & Madden, N. A. Modifying Chapter 1 program improvement guidelines to reward appropriate practices. *Educational Evaluation and Policy Analysis,* 1991, 4, 369-379.

Slavin, R. E. Normal curve equivalent gains: Handle with care! *ERS Bulletin*, 1991, 9, 3-11.

Slavin, R. E., Madden, N. A., Karweit, N. L., Dolan, L. J., & Wasik, B. A. Success for All: Ending reading failure from the beginning. *Language Arts,* 1991, 68, 404-409.

Slavin, R. E. Cooperative learning and group contingencies. *Journal of Behavioral Education*, 1991, 1, 105-115.

Slavin, R. E. Critical commentary on a best-evidence synthesis of effects of passive smoking. *Clinical and Investigative Medicine,* 1991, 14, 486-487.

Slavin, R. E. Early intervention in teaching reading skills prevents failures in higher grades. *School Board News,* 1992, 12 (23), 2, 5.

Gutierrez, R., & Slavin, R. E. Achievement effects of the non-graded elementary school: A best-evidence synthesis. *Review of Educational Research,* 1992, 62, 333-376.

Slavin, R. E. The non-graded elementary school: Great potential, but keep it simple. *Educational Leadership,* 1992, 50 (2), 24-25.

Slavin, R. E., Karweit, N. L., & Wasik, B. A. Preventing early school failure: What works? *Educational Leadership*, 1992/93, 50 (4), 10-18.

Slavin, R. E. Ability grouping in the middle grades: Effects and alternatives. *Elementary School Journal,* 1993, 93, 535-552.

Madden, N. A., Slavin, R. E., Karweit, N. L., Dolan, L., & Wasik, B. Success for All: Longitudinal effects of a schoolwide elementary restructuring program. *American Educational Reseach Journal,* 1993, 30, 123-148.

Wasik, B. A., & Slavin, R. E. Preventing early reading failure with one-to-one tutoring: A best-evidence synthesis. *Reading Research Quarterly,* 1993, 28 (2), 179-200.

Braddock, J. H., & Slavin, R. E. Why ability grouping must end: Achieving excellence and equity in American education. *Journal of Intergroup Relations,* 1993, 20, 51-64.

Mainzer, R. W., Mainzer, K. L., Slavin, R. E., & Lowry, E. What special education teachers should know about cooperative learning. *Teacher Education and Special Education,* 1993, 16 (1), 42-50.

Slavin, R. E. Students differ. So what? *Educational Researcher,* 1993, 22(9), 13-14.

Slavin, R. E., & Braddock, J. H. Ability grouping: On the wrong track. *College Board Review,* 1993, 168, 11-18.

Slavin, R. E. Untracking: The 97% solution. *College Board Review*, 1993, 168, 27, 35.

Slavin, R. E., Madden, N. A., Dolan, L. J., Wasik, B. A., Ross, S., & Smith, L. "Whenever and wherever we choose...:" The replication of Success for All. *Phi Delta Kappan*, 1994, 75 (8), 639-647.

Slavin, R. E. Outcomes-based education is not mastery learning. *Educational Leadership,* 1994, 51 (6), 14-15.

Slavin, R. E. Statewide finance reform: Ensuring educational adequacy for high poverty schools. *Educational Policy,* 1994, 8, 425-434.

Slavin, R. E. Quality, appropriateness, incentive, and time: A model of instructional effectiveness. *International Journal of Educational Research,* 1994, 21, 141-157.

Slavin, R. E. After the victory: Making funding equity make a difference. *Theory into Practice*, 1994, 33, 98-103.

Slavin, R. E., Madden, N. A., Dolan, L. J., & Wasik, B. A. Roots and Wings: Inspiring academic excellence. *Educational Leadership,* 1994, 52 (3), 10-13.

Slavin, R. E. A model of effective instruction. *The Educational Forum,* 1995, 59 (2), 166-176.

Ross, S. M., Smith, L. J., & Slavin, R. E. Increasing the academic success of disadvantaged children: An examination of alternative early intervention programs. *American Educational Research Journal,* 1995, 32(4) 773-800.

Stevens, R. J., & Slavin, R. E. Effects of a cooperative learning approach in reading and writing on handicapped and non-handicapped students. *Elementary School Journal,* 1995, 95 (3), 241-262.

Stevens, R. J., & Slavin, R. E. The cooperative elementary school: Effects on students' achievement, attitudes, and social relations. *American Educational Research Journal,* 1995, 32 (2) 321-351.

Slavin, R. E. Best-evidence synthesis: An intelligent alternative to meta-analysis. *Journal of Clinical Epidemiology,* 1995, 48 (1), 9-18.

Slavin, R. E. Detracking and its detractors: Flawed evidence, flawed values. *Phi Delta Kappan,* 1995, 77 (3), 220-221.

Slavin, R. E. Making money make a difference. *Rethinking Schools,* 1995, 9 (4), 10, 23.

Slavin, R. E., Madden, N. A., Dolan, L.J., Wasik, B. A., Ross, S., Smith, L., & Dianda, M. Success for All: A summary of research. *Journal of Education for Students Placed at Risk,* 1996, 1, 41-76.

Slavin, R. E. Neverstreaming: Preventing learning disabilities. *Educational Leadership,* 1996, 53 (5), 4-7.

Slavin, R. E. Research on cooperative learning and achievement: What we know, what we need to know. *Contemporary Educational Psychology,* 1996, 21 (1), 43-69.

Slavin, R. E. Cooperative learning in middle and secondary schools. *The Clearing House* , 1996, 69 (4), 200-204.

Slavin, R. E. Neverstreaming: Ending learning disabilities before they start. *Support for Learning* (Britain), 1996, 11 (2), 74-76.

Slavin, R. E. Success for All. *Thrust for Educational Leadership,* 1996 (September), 6-9.

Slavin, R. E. A chance to turn schools around. *Baltimore Sun,* December 15, 1996.

Slavin, R. E. Reforming state and federal policies to support adoption of proven practices. *Educational Researcher*, 1996, 25 (9), 4-5.

Fashola, O. S., & Slavin, R. E. Promising programs for elementary and middle schools: Evidence of effectiveness and replicability. *Journal of Education for Students Placed at Risk,* 1997, 2 (3), 251-307.

Slavin, R. E. Design competitions: A proposal for a new federal role in educational research and development. *Educational Researcher*, 1997, 26 (1), 22-28.

Slavin, R. E. Ready...Read! *Baltimore Sun*, May 11, 1997.

Slavin, R. E. How Title I can (still) save America's children. *Education Week,* May 21, 1997.

Slavin, R. E. Design competitions and expert panels: Similar objectives, very different paths. *Educational Researcher*, 1997, 26 (6), 21-22.

Slavin, R.E. Can education reduce social inequity? *Educational Leadership*, 1997/1998, 55 (4), 6-10.

Cooper, R., Slavin, R. E., & Madden, N. A. Success for All: Improving the quality of implementation of whole-school change through the use of a national reform network. In S. Stringfield and A. Datnow (Eds.). *Education and Urban Society,* 1998, 30 (3), 385-408.

Fashola, O. S., & Slavin, R. E. Schoolwide reform models: What works? *Phi Delta Kappan,* 1998, 79 (5), 370-379.

Calderón, M., Hertz-Lazarowitz, R., & Slavin, R. E. Effects of Bilingual Cooperative Integrated Reading and Composition on students transitioning from Spanish to English reading. *Elementary School Journal,* 1998, 99 (2), 153-166.

Fashola, O. S., & Slavin, R. E. Effective dropout prevention and college attendance programs for students placed at risk. *Journal of Education for Students Placed at Risk,* (in press).

Slavin, R.E. Reading by nine: What will it take? *Peabody Journal of Education,* 1998, 73, (3 & 4), 68-80.

Slavin, R.E. Far and wide: Developing and disseminating research-based programs. *American Educator*, 1998, 22 (3), 8-11.

Slavin, R.E. Comprehensive approaches to cooperative learning. *Theory Into Practice,* Spring, 1999, 38 (2), 74-80.

Slavin, R.E. How can funding equity ensure enhanced achievement? *Journal of Education Finance,* 1999, 24 (4), 519-528.

Slavin, R.E. (1999). Yes, control groups are essential in program evaluation: A response to Pogrow. *Educational Researcher, 28* (3), 36-38.

Slavin, R.E., & Madden, N.A. Effects of bilingual and English as a second language adaptations of Success for All on the reading achievement of students acquiring English. *Journal of Education for Students Placed at Risk,* 1999, 4 (4), 393-416.

Slavin, R.E., & Cooper, R. Improving intergroup relations: Lessons learned from cooperative learning programs. *Journal of Social Issues*, 1999, 55 (4), 647-664.

Slavin, R.E. Educational research in the 21st Century: Lessons from the 20th. *Issues in Education,* 1999, 5 (2), 261-266.

Slavin, R.E., & Madden, N.A. Research on achievement outcomes of Success for All: A summary and response to critics. *Phi Delta Kappan,* 2000, 82 (1), 38-40, 59-66.

Slavin, R.E., & Madden, N.A. Roots & Wings: Effects of whole school reform on student achievement. *Journal of Education for Students Placed at Risk,* 2000, 5 (1 & 2), 109-136.

Slavin, R.E. Putting the school back in school reform. *Educational Leadership*, 2000/2001, 58 (4), 22-27.

Slavin, R.E. Expecting excellence. *American School Board Journal,* 2000, 188 (2), 22-25.

Slavin, R.E. Show me the evidence. *American School Board Journal,* 2001, 188 (3), 26-29.

Slavin, R.E. Reauthorizing Title I: A comprehensive vision. *Title I Monitor,* 2001, 6 (2), 14-16.

Hurley, E., Chamberlain, A., Slavin, R.E., & Madden, N.A. Effects of Success for All on TAAS Reading: A Texas statewide evaluation. *Phi Delta Kappan*, 2001, 82 (10), 750-756.

Slavin, R.E. Response. The Facts About Comprehensive School Reform. *Educational Leadership*, 2001, 59 (1), 84-85.

Slavin, R.E. (2002). Mounting evidence supports the achievement effects of Success for All. *Phi Delta Kappan, 83* (6), 469-471, 480.

Slavin, R.E. (2001). Big change questions: Does rigorous research knowledge provide a better base for improving student achievement than teachers' craft knowledge? *Journal of Educational Change, 2* (4), 347-348.

Slavin, R.E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher,* 31 (7), 15-21.

Slavin, R.E. (2003). A reader's guide to scientifically based research. *Educational Leadership, 60* (5), 12-16.

Slavin, R.E. (2003, March 5). Converging reforms: Change schools? Change districts? How the two approaches can work together. *Education Week*, 64, 44-45.

Slavin, R.E. (2004). Built to last: Long-term maintenance of Success for All. *Remedial and Special Education, 25* (1), 61-67.

Slavin, R.E. (2004). Education research can and must address "what works" questions. *Educational Researcher, 33* (1), 27-28.

Slavin, R.E. & Cheung, A. (2004). How do English language learners learn to read? *Educational Leadership, 61* (6), 52-57.

Chambers, B., Slavin, R.E., Madden, N.A., Cheung, A., & Gifford, R. (2004). *Effects of Success for All with embedded video on the beginning reading achievement of Hispanic children.* Manuscript submitted for publication.

Borman, G., Slavin, R.E., Cheung, A., Chamberlain, A., Madden, N.A., & Chambers, B. (2005). Success for All: First year results from the national randomized field trial. *Educational Evaluation and Policy Analysis, 27* (1), 1-22.

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A., Madden, N. A., & Chambers, B. (2005). *The National Randomized Field Trial of Success for All: Second-Year outcomes.* Manuscript submitted for publication.

Slavin, R.E. (2005). Evidence-based reform in education: Promise and pitfalls. *Mid-Western Educational Researcher, 18* (1), 8-13.

Slavin, R.E., Daniels, C., & Madden, N.A. (2005). The Success for All Middle School: Adding content to middle grades reform. *Middle School Journal*, 36 (5), 4-8.

Cheung, A., & Slavin, R.E. (in press). Effective reading programs for English language learners and other language minority students. *Bilingual Research Journal.*

Chambers, B., Cheung, A., Madden, N. A., Slavin, R. E., & Gifford, G. (in press). Achievement effects of embedded multimedia in a Success for All reading program. *Journal of Educational Psychology*.

Slavin, R. E., & Cheung, A. (in press). Effective reading programs for English language learners: A best evidence synthesis. *Review of Educational Research*.

Book Chapters

Slavin, R. E., & DeVries, D. L. Learning in teams. In H. Walberg (Ed.), *Educational environments and effects*. Berkeley, Calif.: McCutchan, 1979.

Slavin, R. E. Student Team Learning: A manual for teachers. In S. Sharan, P. Hare, C. Webb, and R. Hertz-Lazarowitz (Eds.), *Cooperation in education*. Provo, Utah: Brigham Young University Press, 1980 .

Slavin, R. E. Cooperative learning and desegregation. In W. Hawley (Ed.), *Effective desegregated schools*, Beverly Hills, CA: Sage, 1981.

Slavin, R. E. Non-cognitive outcomes of cooperative learning. In J. Levine and M. Wang (Eds.), Teacher and student perceptions: *Implications for learning*. Hillsdale, NJ: Erlbaum, 1982.

Slavin, R. E. A case study of psychological research affecting classroom practice: Student Team Learning. In W. Doyle and T. Good (Eds.), Focus on teaching: *Readings from the Elementary School Journal*. Chicago: University of Chicago Press, 1982.

Slavin, R. E., & Hansell, S. Cooperative learning and inter-group relations: Contact theory in the classroom. In J. Epstein and N. Karweit (Eds.), *Friends in school*. New York: Academic Press, 1983.

Slavin, R. E. Gruppenrallye (Student Teams-Achievement Divisions). In G. L. Huber, S. Rotering-Steinberg, and D. Wahl (Eds.), *Kooperatives Lernen*. Weinheim, West Germany: Beltz, 1984.

Slavin, R. E. Team Assisted Individualization. In R. E. Slavin, S. Sharan, S. Kagan, R. Hertz-Lazarowitz, C. Webb, and R. Schmuck (Eds.), Learning to cooperate, cooperating to learn. New York: Plenum, 1985. Reprinted in Barber, L. (1986). *Adapting instruction to individual needs: An eclectic approach*. Bloomington, IN: Phi Delta Kappa .

Slavin, R. E. An introduction to cooperative learning research. In R. E. Slavin, S. Sharan, S. Kagan, R. Hertz-Lazarowitz, C. Webb, and R. Schmuck (Eds.), *Learning to cooperate, cooperating to learn*. New York: Plenum, 1985.

Slavin, R. E. Small-group instruction. In T. Husen and T. N. Postlewaite (Eds.), *International Encyclopedia of Education*. Oxford: Pergamon, 1985. Updated version published in International Encyclopedia of Teaching and Teacher Education. Oxford: Pergamon, 1986.

Slavin, R. E. Team Assisted Individualization: A cooperative learning solution for adaptive instruction in mathematics. In M. Wang and H. Walberg (Eds.), *Adapting instruction to individual differences*. Berkeley, CA: McCutchan, 1985.

Slavin, R. E. Cooperative learning: Engineering social psychology in the classroom. In R. Feldman (Ed.), *The social psychology of education: Current research and theory*. Cambridge: Cambridge University Press, 1986.

Slavin, R. E. Cooperative learning and the education of black students. In D. Strickland and E. Cooper (Eds.), *Educating black children: America's challenge*. Washington D.C.: Howard University Press, 1987.

Slavin, R. E. A theory of school and classroom instruction. In R. E. Slavin (Ed.), *School and classroom organization*. Hillsdale, NJ: Erlbaum, 1989.

Slavin, R. E. Grouping for instruction in the elementary school. In R. E. Slavin (Ed.), *School and classroom organization*. Hillsdale, NJ: Erlbaum, 1989.

Slavin, R. E. Achievement effects of group-based mastery learning. In R. E. Slavin (Ed.), *School and classroom organization*. Hillsdale, NJ: Erlbaum, 1989.

Slavin, R. E. Cooperative learning and student achievement. In R. E. Slavin (Ed.), *School and classroom organization*. Hillsdale, NJ: Erlbaum, 1989.

Slavin, R. E. Achievement effects of substantial reductions in class size. In R. E. Slavin (Ed.), *School and classroom organization*. Hillsdale, NJ: Erlbaum, 1989.

Slavin, R. E. A cooperative learning approach to reading in content areas: Jigsaw Teaching. In D. Lapp, J. Flood, and N. Farnan, (Eds.), *Instructional theory and practice for content area reading and learning*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

Slavin, R. E. Student team learning in mathematics. In N. Davidson (Ed.), *Small group cooperative learning in mathematics*. Boston: Addison Wesley, 1989.

Slavin, R. E. Cooperative learning and achievement: Six theoretical perspectives. In M. L. Maehr and C. Ames (Eds.), *Advances in motivation and achievement*. Greenwich, CT: JAI Press, 1989.

Slavin, R. E. Students at risk of school failure: The problem and its dimensions. In R. E. Slavin, N. L. Karweit, and N. A. Madden (Eds.), *Effective programs for students at risk.* Boston, MA: Allyn and Bacon, 1989.

Slavin, R. E., & Madden, N. A. Effective classroom programs for students at risk. In R. E. Slavin, N. L. Karweit, and N. A. Madden (Eds.), *Effective programs for students at risk.* Boston, MA: Allyn and Bacon, 1989.

Madden, N. A., & Slavin, R. E. Effective pullout programs for students at risk. In R. E. Slavin, N. L. Karweit, and N. A. Madden (Eds.), *Effective programs for students at risk.* Boston, MA: Allyn and Bacon, 1989.

Slavin, R. E., Madden, N. A., & Karweit, N. L. Effective programs for students at risk: Conclusions for practice and policy. In R. E. Slavin, N. L. Karweit, and N. A. Madden (Eds.), *Effective programs for students at risk.* Boston, MA: Allyn and Bacon, 1989.

Stainback, S., Stainback, W., & Slavin, R. E. Classroom organization for diversity among students. In S. Stainback, W. Stainback, and M. Forest (Eds.), *Educating all students in the mainstream of regular education.* Baltimore: Brookes, 1989.

Slavin, R. E. Comprehensive cooperative learning models: Embedding cooperative learning in the curriculum and the school. In S. Sharan (Ed.), *Cooperative learning: Theory and research.* New York: Praeger, 1990.

Slavin, R. E., & Stevens, R. J. Cooperative learning and mainstreaming. In J. W. Lloyd, N. N. Singh, and A. C. Repp (Eds.), *The regular education initiative: Alternative perspectives on concepts, issues, and models* (pp. 177-192). Sycamore, IL: Sycamore, 1991.

Slavin, R. E. When and why does cooperative learning increase achievement? Theoretical and empirical perspectives. In R. Hertz-Lazarowitz and N. Miller (Eds.), *Interaction in cooperative groups: The theoretical anatomy of group learning.* New York: Cambridge University Press, 1992.

Slavin, R. E. Grouping. In L. R. Williams & D. P. Fromberg (Eds.), *Encyclopedia of early childhood education.* New York: Garland, 1992..

Slavin, R. E. Co-operative learning. In C. Rogers and R. Kutnick (Eds.) *The social psychology of the primary school.* London: Routledge, 1992.

Slavin, R. E. Cooperative learning in social studies: Balancing the social and the studies. In R. J. Stahl and R. L. VanSickle (Ed.), *Cooperative learning in the social studies classroom.* Washington, D.C.: National Council for the Social Studies, 1992.

Slavin, R. E., Madden, N. A., Shaw, A. H., Mainzer, K. L., & Donnelly, M. C. Success for All: Three case studies of comprehensive restructuring of urban elementary schools. In J. Murphy and P Hallinger (Eds.), *Restructuring schooling: Learning from ongoing efforts.* Newbury Park, CA: Corwin, 1993.

Slavin, R. E. Preventing early school failure: The challenge and the opportunity. In R. E. Slavin, N. L. Karweit, and B. A. Wasik (Eds.), *Preventing early school failure*. Boston: Allyn and Bacon, 1994.

Slavin, R. E. School and classroom organization in beginning reading: Class size, aides, and instructional grouping. In R. E. Slavin, N. L. Karweit, and B. A. Wasik (Eds.), *Preventing early school failure*. Boston: Allyn and Bacon, 1994.

Slavin, R. E., Madden, N. A., Karweit, N. L., Dolan, L. J., & Wasik, B. A. Success for All: A comprehensive approach to prevention and early intervention. In R. E. Slavin, N. L. Karweit, and B. A. Wasik (Eds.), *Preventing early school failure*. Boston: Allyn and Bacon, 1994.

Slavin, R. E. Preventing early school failure: Implications for policy and practice. In R. E. Slavin, N. L. Karweit, and B. A. Wasik (Eds.), *Preventing early school failure*. Boston: Allyn and Bacon, 1994.

Slavin, R. E., Madden, N. A., & Stevens, R. J. Cooperative integrated reading and composition: Applications for the language arts classroom. In R. J. Stahl (Ed.), *Cooperative learning in language arts: A handbook for teachers.* Boston: Addison-Wesley, 1994.

Slavin, R. E. and Braddock, J. H. Ability grouping: On the wrong track. In Goodlad, J. I., & Keating, P. (Eds.), *Access to knowledge: The continuing agenda for our nation's schools* New York: The College Board, 1994.

Slavin, R. E. Cooperative learning and intergroup relations. In J. A. Banks & C. A. Banks (Eds.), *Handbook of research on multicultural education..* New York: Macmillan, 1995.

Slavin, R. E. Cooperative learning and student diversity. In R. Ben-Ari and Y. Rich (Eds.), *Understanding and treating diversity in education: An international perspective.* New York: Macmillan, 1995.

Slavin, R. E., & Madden, N. A. Success for All: Creating schools and classrooms in which all children can read. In J. Oakes (Ed.), *Creating new educational communities.* Chicago: University of Chicago Press, 1995.

Slavin, R. E. Teaching at-risk students: What works? In A. C. Ornstein (Ed.), *Teaching: Theory into practice.* Boston: Allyn & Bacon, 1995.

Slavin, R. E. Cooperative learning and student diversity. In J. A. Banks & C. A. M. Banks (Eds.), *Handbook of research on multicultural education* . New York: Macmillan, 1995.

Slavin, R. E. Enhancing intergroup relations in schools: Cooperative learning and other strategies. In W. D. Hawley and A. W. Jackson (Eds.), *Toward a common destiny:* San Francisco: Jossey-Bass, 1995.

Braddock, J.H., & Slavin, R.E. Why ability grouping must end: Achieving excellence and equity in American education. In H. Pool & J.A. Page (Eds.), *Beyond tracking*. Bloomington, IN: Phi Delta Kappa, 1995.

Gutiérrez, R., & Slavin, R. E. Research in nongraded elementary schools: Simple is beautiful. In R. Fogarty (Ed.), *Think about...Multiage classrooms: An anthology of original essays*. Illinois: IRI/Skylight Training and Publishing, 1995.

Slavin, R. E., Madden, N. A., & Wasik, B. A. Roots and Wings. In S. Stringfield, S. Ross, & L. Smith (Eds.), *Bold Plans for Educational Reform: The New American Schools*. Hillsdale, NJ: Erlbaum, 1996.

Madden, N. A., Slavin, R. E., Wasik, B. A., & Dolan, L. J. Reading, writing, and language arts in Success for All. In S. Stahl and D. Hayes (Eds.), *Instructional models in reading*. Mahwah, NJ: Erlbaum, 1997.

Slavin, R.E. Co-operative learning among students. In D. Stern & G. Huber (Eds.), *Active Learning for students and teachers: Reports from eight countries*. Paris: Organization for Economic Co-Operation and Development, 1997.

Slavin, R.E. Cooperative learning and student diversity. In R. Ben-Ari & Y. Rich (Eds.), *Enhancing education in heterogeneous Schools*. Ramat-Gan: Bar-Ilan University Press, 1997.

Slavin, R. E. Sand, bricks, and seeds: School change strategies and readiness for reform. In A. Hargreaves, A. Liberman, M. Fullan, & D. Hopkins (Eds.). *International handbook of educational change*. Dardrecht, The Netherlands: Kluwer, 1998.

Slavin, R.E. Success for All: Policy consequences of a replicable schoolwide reform. In G.J. Cizek (Ed.), *Handbook of educational policy*. Orlando, FL: Academic Press, 1998.

Slavin, R.E. The pendulum revisited: Faddism in education and its alternatives. In G.J. Cizek (Ed.), *Handbook of educational policy*. Orlando, FL: Academic Press, 1999.

Slavin, R. E. (1999). Educating young students at risk of school failure: Research, practice and policy. In Robert J. Stevens (Ed.), *Teaching in American schools*. Upper Saddle River, NJ: Merrill.

Slavin, R.E., & Madden, N.A. Success for All: Effects of prevention and early intervention on elementary students' reading. In Lotty Eldering and Paul P.M. Leseman (Eds.), *Effective Early Education*. New York, NY: Falmer Press, 1999.

Slavin, R.E., & Madden, N.A. (1999). Roots & Wings: A comprehensive approach to elementary school reform. In J. Block, S. Everson, & T. Guskey (Eds.), *Comprehensive school reform: A program perspective*. Dubuque, IA: Kendall/Hunt.

Slavin, R.E., & Madden, N.A. (2001). Effects of bilingual and English as a second language adaptations of Success for All on the reading achievement of students acquiring English. In R.E. Slavin & M. Calderón (Eds.), *Effective programs for Latino students.* Mahwah, NJ: Erlbaum.

Fashola, O.S., & Slavin, R.E. (2001). Effective dropout prevention and college attendance programs for Latino students. In R.E. Slavin and M.Calderón, (Eds.), *Effective programs for Latino students*. Mahwah, NJ: Erlbaum

Fashola, O.S., Slavin, R.E., Calderón, M., & Durán, R. (2001). Effective programs for Latino students in elementary and middle schools. In R.E. Slavin and M.Calderón, (Eds.), *Effective programs for Latino students*. Mahwah, NJ: Erlbaum.

Slavin, R.E., Hurley, E.A., & Chamberlain, A.M. (2001). Cooperative learning in schools. In N.J. Smelser and Paul B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 2756-2761). Oxford, England: Pergamon.

Slavin, R.E., and Madden, N.A. (2001). Summary of research on Success for All and Roots & Wings. In R.E. Slavin & N.A. Madden (Eds.), *Success for All: Research and reform in elementary education.* Mahwah, NJ: Erlbaum.

Slavin, R.E., and Madden, N.A. (2001). Disseminating Success for All: Lessons for policy and practice. In R.E. Slavin & N.A. Madden (Eds.), *Success for All: Research and reform in elementary education.* Mahwah, NJ: Erlbaum.

Slavin, R.E. (2001). How Title I can become the engine of reform in America's schools. In Borman, G.D., Stringfield, S.C., & Slavin, R.E. (Eds.), *Title I: Compensatory education at the crossroads.* Mahwah, NJ: Erlbaum.

Slavin, R.E. (2001). Success for All: failure prevention and early intervention. In Sornson, B. (Ed.), *Preventing early learning failure.* Alexandria, VA: ASCD.

Cooper, R., & Slavin, R.E. (2001). Cooperative learning programs and multicultural education: Improving intergroup relations. In Salili, F. & Hoosain, R. (Eds.), *Multicultural education: Issues, policies, and practices.*

Slavin, R.E., Hurley, E.A., & Chamberlain, A.M. (2001). Cooperative learning in schools. In N.J. Smelser and Paul B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 2756-2761). Oxford, England: Pergamon.

Slavin, R.E. (2002). Cooperative learning in elementary and secondary schools. In Levinson, D.L., Cookson, Jr., P.W., & Sadovnik, A.R. (Eds.), *Education and sociology.* New York, NY: Routledge Falmer.

Slavin, R.E. (2002). Cooperative learning. In Salkind, N.J. (Ed.), *Child Development.* New York, NY: Macmillan Library Reference USA.

Slavin, R.E. (2002). The intentional school: Effective elementary education for all children. In Stringfield, S., & Land, D. (Eds.), *Educating at-risk students*. Chicago, IL: University of Chicago.

Slavin, R.E. & Madden, N.A. (2002). "Success for All" and African-American and Latino student achievement. In Chubb, J. & Loveless, T. (Eds.), *Bridging the Achievement Gap*. Washington, DC: Brookings Institution Press.

Slavin, R.E., Hurley, E.A., & Chamberlain, A.M. (2003). Cooperative learning and achievement: Theory and research. In W.M. Reynolds & G.E. Miller (Eds.), *Handbook of Psychology*, Volume 7 (pp. 177-198). Hoboken, NJ: Wiley.

Slavin, R.E., & Cheung, A. (2004). Effective early reading programs for English Language Learners. In O. Saracho & B. Spodek (Eds.), *Contemporary Perspectives on Language Policy and Literacy Instruction in Early Childhood Education*. Greenwich, CT: Information Age Publishing.

Educational Evaluation and Policy Analysis
Spring 2005, Vol. 27, No. 1, pp. 1–22

Success for All: First-Year Results From the National Randomized Field Trial

Geoffrey D. Borman
University of Wisconsin, Madison

Robert E. Slavin
Johns Hopkins University

Alan Cheung, Anne M. Chamberlain, Nancy A. Madden, and Bette Chambers
Success for All Foundation

This article reports first-year achievement outcomes of a national randomized evaluation of Success for All, a comprehensive reading reform model. Forty-one schools were recruited for the study and were randomly assigned to implement Success for All or control methods. No statistically significant differences between experimental and control groups were found in regard to pretests or demographic characteristics. Hierarchical linear model analyses revealed a statistically significant school-level effect of assignment to Success for All of nearly one quarter of a standard deviation—or more than 2 months of additional learning—on individual Word Attack test scores, but there were no school-level differences on the three other posttest measures assessed. These results are similar to those of earlier matched experiments and correspond with the Success for All program theory.

Keywords: *educational policy, experimental design, school reform*

RECENT national policy has focused on replicating and scaling up research-proven educational programs. In several important ways, Success for All has provided a compelling model of the type of program such policies aim to support. Of 33 comprehensive school reform programs reviewed in a recent meta-analysis, it was one of only 3 that exhibited positive and statistically significant achievement effects across a large number of rigorous quasi-experimental studies (Borman, Hewes, Overman, & Brown, 2003). Even during the original development of the program's core components, which are summarized in the Appendix, the developers conceived of them entirely from their own research and the research of others on "what works" in education reform (Slavin & Madden, 2001). Since the program's origination in the late 1980s, the Success for All developers have built the capacity to disseminate and maintain implementations of their programs at a national scale, currently in more than 1,400 schools in 47 states. First based at Johns Hopkins University and now at the nonprofit Success for All Foundation, Success for All is one of only a few programs capable of adding more than a hundred schools to its network each year.

Although Success for All has accumulated a substantial body of quasi-experimental evidence, there have not been any true randomized experiments of the program's effects. This lack of experimental evidence was not perceived by policymakers and researchers as a problem in the past, but in the current national policy environment such an omission is beginning to have increasingly important consequences. For the first time, Congress and other education policymakers are requiring strong evidence of effectiveness among programs seeking certain types of funding. In 1997, this trend began with the Comprehensive School Reform Demonstration (CSRD) program, which provided grants to schools to adopt "proven, comprehensive reform models." More recently, the Reading Excellence Act (REA) and its successor, Reading First, have required that grant funds be used to help schools adopt those programs that incorporate "scientifically based principles" of reading instruction.

Finally, the 2002 No Child Left Behind Act reauthorizing Title I, the federal government's largest single investment in the nation's elementary and secondary schools, referred to "scientifically based research" more than 100 times (Olson & Viadero, 2002). This law has placed a premium on randomized experiments designed to develop and assess new and innovative practices, as the following excerpt suggests: "The Secretary shall evaluate the demonstration projects supported under this title, using rigorous methodological designs and techniques, including control groups and random assignment, to the extent feasible, to produce reliable evidence of effectiveness" (No Child Left Behind Act, 2002).

Policies of this kind greatly raise the stakes for research in education. Although in practice CSRD, REA, Reading First, and Title I grants have not held schools to high standards of evidence, the movement to base educational policies on research must ultimately depend on building a stronger research base for replicable programs. Success for All is a logical program to be held to a new and higher standard of evidence because of the substantial foundation of research it already has built from quasi-experimental studies and because it has been the most widely disseminated program funded by the CSRD legislation.

To set a higher standard of evidence for policy-relevant research in education, and to address some of the past criticisms of research on Success for All, the Success for All Foundation applied for and received funding to carry out the first randomized evaluation of the program. This grant was one of six awarded by the Office of Educational Research and Improvement (OERI) for the study of comprehensive school reform models. The other five projects are investigating various combinations of comprehensive school reform models, including Success for All. The Success for All Foundation grant, however, is the only one to include a randomized experiment.

The OERI grant was originally proposed as a first-party evaluation involving strong, independent review by an oversight committee composed of distinguished scholars not previously connected to Success for All.[1] However, in response to clarification requests from OERI, the Success for All Foundation agreed to have the data collected by a third party and ultimately contracted with the National Opinion Research Center at the University of Chicago.

Here we present the first-year findings from this national randomized field trial. We begin by examining the internal and external validity of the experiment. Specifically, did randomization yield treatment and control schools with comparable characteristics, and did sample and data attrition over the first year jeopardize this baseline comparability? Regarding the external validity of the experiment, how representative of the population of Success for All schools was this sample of schools, and did sample and data attrition have an impact on the generalizability of the results? Finally, we examine the achievement outcomes for the kindergarten and first-grade cohorts that were tested from fall to spring during the first year. The two-level hierarchical linear models that we formulated were intended to provide an answer to the question of whether there are school-level effects of treatment assignment on four particular measures of literacy achievement.

Success for All Research Base

More than 40 separate studies of Success for All have been conducted by investigators across a large number of U.S. cities and states. The cumulative evidence from these studies shows positive effects of Success for All on a variety of measures of student achievement, as well as on assignments to special education, retentions, and other outcomes (Borman et al., 2003; Slavin & Madden, 2001). Nevertheless, the evidence base supporting Success for All does have certain inherent weaknesses, and it has been the subject of criticism by some educational researchers, including Pogrow (2000) and Walberg and Greenberg (1999).

First, the typical research design in previous evaluations of the program has been a quasi-experimental, untreated control group design in which schools using Success for All have been compared over time with schools not using Success for All matched on demographic characteristics, previous achievement levels, and other factors. Although this is a sound design, all such quasi-experiments involve limitations (Shadish, Cook, & Campbell, 2002). Most important, they leave open the possibility of selection artifacts explaining some or all of the differences observed between the Success for All and control groups.

With such a design, there always may be some systematic reason that the experimental group implemented the program while the comparison group did not. Schools whose staffs have expressed interest in Success for All and achieved the required 80% majority vote to adopt it may have greater motivation and interest in improving their schools than control schools whose staffs have not sought out the program. As indicated by the 80% agreement among staff, the former schools may have strong cohesion among teachers or may have better leaders. Perhaps the experimental schools have better funding or fewer demands on their resources or energies. Alternatively, perhaps the experimental schools are experiencing greater difficulties and have more need for change. These potential artifacts can make it difficult to know whether improvements in schools are the product of characteristics related to selection of the Success for All program or of the components of the program. Most studies of Success for All have been well-designed matched experiments that have minimized selection bias—for example, by designating control schools in advance and by avoiding the use of control schools that have rejected the program—but selection bias cannot be ruled out in the absence of random assignment.

Second, there has been criticism of studies of Success for All that have been conducted by the developers themselves. Borman et al. (2003) found that studies of comprehensive school reform programs conducted by the developers tend to report higher estimates of the programs' effects on achievement outcomes. After statistical adjustment for other methodological differences across the studies, the magnitude of this difference is between one sixth and one seventh of a standard deviation relative to studies completed by researchers other than the developers. As Borman and his colleagues noted, one of the most likely explanations for this difference is that when developers are more actively involved in the study of their models, they are also more likely to be actively involved in ensuring that they are studying a high-quality implementation. In this respect, many of these studies may represent what Cronbach et al. (1980) termed the "superrealization" stage of program development. Before broad field trials are initiated, interventions are often studied under optimal conditions as assessments of what the program can accomplish at its best. However, the extent to which the developers' studies and results may generalize across broader implementations of their school reform models is of some concern.

Critics of small-scale quasi-experiments also may argue that the use of small samples of program schools and matched control schools that are, essentially, handpicked by the developer can introduce particularly advantageous comparisons that overstate the impact of the school reform program. In a large randomized study, the process of choosing program and control schools is not left to the developer but is determined by, for instance, the arbitrary flip of a coin. Handpicking schools and making comparisons that "look good" for the program are not possible within a randomized design, especially when the process of randomization is the responsibility of a neutral third party. As a result of potential limitations such as those just described, Success for All—like any other educational program—could benefit from greater third-party involvement in broader field trials involving true experimental designs.

Role of Randomized Field Trials in Education Policy

Recently, educational research has come under attack for its dearth of randomized experiments (Borman, 2003; Boruch, de Moya, & Snyder, 2002; Cook, 2002; Cook & Payne, 2002). Indeed, the number of experiments conducted to inform other types of social policy and other disciplines within the social sciences, such as psychology and criminology, is increasing at a much faster rate than in education (Boruch et al., 2002). The information gained in these other social arenas is proving to be very useful for identifying promising programs and avoiding ineffective and harmful ones (Petrosino, Turpin-Petrosino, & Buehler, 2003). Evidence from across the social sciences also indicates that experiments are more efficient than other techniques in that they provide more consistent results across replications (Bloom, Michaelopoulous, Hill, & Lei, 2002; Cook, 2002; Lipsey & Wilson, 1993) and are less likely to produce biased results (Glazerman, Ley, & Myers, 2002).

Although few randomized experiments have been conducted in education, those that have been carried out have been very influential. For instance, the randomized longitudinal Tennessee Class Size Study (Finn & Achilles, 1999) led directly to massive class-size reduction initiatives in several states, notably California, and to the Clinton administration's national class-size initiative. The randomized, longitudinal evaluation of the Perry Preschool (Schweinhart, Barnes, & Weikart, 1993) led to substantial expansion of the federal Head Start program and to publicly funded preschool programs in many states and localities. Finally, the promising results from the Abecedarian Project (Ramey & Campbell, 1984), which randomly assigned mothers and their infants to a highly intensive educational child-care program beginning shortly after birth, inspired the U.S. Congress to develop and then scale up the Early Head Start program.

Beyond the scientific importance of random assignment, the political importance of rigorous evidence continues to grow. If federal education legislation continues the trend toward linking funding of education programs to evidence of effectiveness, the consequences could be revolutionary. If education reform can be based on rigorous research, then genuine progress in educational practice becomes possible, as in medicine, agriculture, technology, and other parts of modern economies that long ago accepted the idea that progress must be based on rigorous research and development. In place of the famous pendulum swinging, for example, from phonics to whole language or tracking to untracking, there would be widely accepted findings providing a solid basis for policy and practice.

Imagine, for example, that schools implementing Title I programs, bilingual programs, special education programs, or dropout prevention programs were encouraged or required to adopt programs and practices that had evidence of effectiveness from high-quality experimental evaluations. This transformation of the education reform landscape will not take place, however, unless there are convincing demonstrations that replicable programs can accelerate student achievement. These demonstrations must be beyond reproach and must involve the most rigorous evaluation methods known. In other words, they must involve random assignment, extensive measurement of implementations and outcomes, and longitudinal designs on a sufficiently large scale to ensure both statistical power and generalizability.

Success for All National Randomized Field Trial

This article presents the first-year findings from a longitudinal study that should add enormously to knowledge about the effectiveness of Success for All in increasing the achievement and school success of students placed at risk. First, it is the initial randomized study of Success for All, virtually eliminating selection bias as an alternative explanation for any of the results observed. Second, it is the largest study ever to compare Success for All and control schools, enabling the use of appropriate statistical methods, especially hierarchical linear modeling, with adequate statistical power to detect true differences. Third, the qualitative data we will examine in future reports will help us measure implementation and relate it to important outcomes. These analyses will allow us to test the program theory and to examine more clearly how and why implementation of the various Success for All components may affect achievement.

Taken together, the project's interrelated activities will provide data of unprecedented richness, detail, and methodological rigor to inform educators about the effects of Success for All, the reasons for these effects, and the conditions under which the effects are most likely to be obtained. In the sections to follow, we present analyses of the first-year achievement outcomes of the project and assess the strength of the foundation that has been laid for this longitudinal effort.

Method

Sample Selection

Recruitment of schools for the randomized study began in November 2000. From the outset, there were problems in providing sufficient incentives to induce school leaders to allow their schools to be assigned at random to experimental or control conditions. Initial efforts focused on reducing the cost to schools of implementing

Success for All, which would ordinarily require schools to allocate about $75,000 in the first year, $35,000 in the second year, and $25,000 in the third year. At first, schools willing to be assigned at random to Success for All or control conditions were offered a $20,000 discount on first-year program costs. By spring 2001, this incentive had been increased to offer schools in either assigned condition a one-time payment of $30,000 in exchange for participation in the study. Control schools could use the incentive in whatever way they wished and were allowed to implement any innovation other than Success for All. However, this incentive did not attract an adequate number of schools. Six schools were recruited in this manner by summer 2001, and these schools were randomly assigned to experimental and control conditions. This number was far from sufficient.

By late spring 2002, a satisfactory (though expensive) incentive structure was in place. Schools willing to participate were assigned at random to use Success for All either in Grades K–2 or in Grades 3–5, at no cost. In this way, all schools would receive at least part of the program, and they did not have to contribute significant amounts of money in a time of tightened budgets. This incentive was sufficient, and we were able to recruit a total of 41 schools, including the 6 schools from the previous year. Grades K–2 in the schools assigned to the 3–5 condition served as the controls for the schools assigned to the K–2 condition, and vice versa.

This design, which included both treatment and control conditions within each school, involved advantages as well as disadvantages. Clearly, the design was important for successful recruitment of schools, and it produced valid counterfactuals for the experimental groups that represented what would have occurred had the experiment not taken place. The danger in the design was that aspects of the treatment grades might affect instruction in the nontreatment grades. Observations intended to determine Success for All treatment fidelity, described subsequently, failed to document contamination of this kind; to the extent that it may have taken place, however, such contamination would have worked against finding differences between the experimental and control groups. Similarly, having the two treatments in the same school may have reduced the estimated effectiveness of school-level aspects of the treatment, such as family support, because control students (in addition to treatment students) could have come forward to take advantage of these services. Thus, any limitations of the design would serve to underestimate, rather than overstate, the treatment effects. Still, treatment fidelity observations suggested that the materials and instructional procedures employed in the Success for All and control grades were distinct from each other and that no control students benefited directly from school-level Success for All services.

All schools (and their districts) had to agree to allow for individual and group testing of their children, to allow observers and interviewers access to the school, and to make available (in coded form, to maintain confidentiality) routinely collected student data such as attendance, disciplinary referrals, special education placements, retentions, and so on. In addition, schools had to agree to allow data collection for 3 years and to remain in the same treatment condition for all 3 years. Schools that agreed to these conditions were randomly assigned by members of the study oversight committee to experimental or control conditions.

A description of the final sample of 41 schools recruited for the study is provided in Table 1. The 6 pilot schools recruited in 2001–2002 either are implementing the entire Success for All program or are entirely control schools. That is, a pilot Success for All school is analyzed as a K–2 Success for All school as well as, in the future, a 3–5 Success for All school. A pilot control school is analyzed as such in both categories. For this reason, the controls for the K–2 study include 17 schools implementing Success for All in Grades 3–5 and the 3 pilot control schools. The total sample of children in kindergarten and first grade in fall 2002 is considered the main longitudinal sample. Students who were in kindergarten or first grade in fall 2001 in the 6 pilot schools are also included.

As is clear from Table 1, the treatment and control samples are reasonably well matched on baseline demographics. The sample is concentrated in the urban Midwest (Chicago, St. Louis, and Indianapolis) and the rural and small-town South, although there are several exceptions. Overall, the students in the sample are very disadvantaged, with a few rural exceptions. About 76% of the students qualify for free lunch, similar to the 80% free-lunch participation rate in the nationwide population of Success for All schools.

TABLE 1
Schools Participating in Randomized Evaluation of Success for All, Grouped by Assignment

Assignment and school	District	State	Enrollment	White (%)	African American (%)	Hispanic (%)	Female (%)	ESL (%)	Special education (%)	Free lunch (%)
Treatment										
Northwood[a]	Mooresville	IN	424	94.70	0.00	0.00	49.50	0.00	4.00	26.00
Jefferson[a]	Midland	OH	315	98.50	2.00	0.00	49.00	0.00	16.00	29.00
Bertha S. Sternberger	Guilford	NC	381	62.00	30.00	5.00	50.00	4.00	20.00	30.00
Pleasant Garden	Guilford	NC	698	76.00	13.70	4.56	49.50	3.50	16.00	33.00
Waveland	S. Montgomery	IN	152	99.00	0.00	1.00	46.00	0.00	15.00	34.00
James Y. Joyner	Guilford	NC	454	41.80	42.90	5.00	48.50	0.05	38.00	35.00
Laurel Valley	Ligonier Valley	PA	409	99.98	0.01	0.01	47.00	0.00	9.00	51.00
Wood	Tempe	AZ	642	19.60	9.60	40.10	49.70	37.50	11.00	51.70
Cesar Chavez	Norwalk	CA	589	6.00	3.00	88.00	47.00	39.00	5.00	71.00
Haven[a]	Savannah	GA	373	0.00	99.00	0.00	65.00	0.00	5.00	85.00
Brian Piccolo	Chicago	IL	1069	0.10	79.50	20.10	48.00	11.40	11.60	85.60
Robert H. Lawrence	Chicago	IL	635	0.00	98.60	0.02	60.00	0.00	11.30	89.00
Harriett B. Stowe	Indianapolis	IN	174	28.71	25.74	43.56	38.00	40.59	17.80	90.59
Linden	Linden	AL	291	1.00	97.00	1.00	49.00	0.00	15.90	91.00
Lafayette	St. Louis	MO	297	13.20	72.80	9.40	49.70	25.00	12.00	94.00
Benjamin E. Mays	Chicago	IL	263	0.00	95.00	5.00	49.00	0.00	10.00	95.00
Paramount Jr.	Greene	AL	493	0.00	100.00	0.00	41.00	0.00	11.00	97.00
Farragut	St. Louis	MO	350	0.00	100.00	0.00	44.00	0.00	4.00	98.00
Gundlach	St. Louis	MO	365	0.00	100.00	0.00	48.00	0.00	4.40	99.00
Cook	St. Louis	MO	335	0.00	100.00	0.00	45.00	0.00	15.00	100.00
Earl Nash	Noxubee	MS	509	1.00	98.00	1.00	51.00	0.20	6.50	100.00
Treatment school means			439	30.60	55.60	10.60	48.76	7.68	12.31	70.70
Control										
Newby[a]	Mooresville	IN	252	98.0[illegible]	0.00	0.00	53.00	0.00	10.00	[illegible]
Jamestown	Guilford	NC	516	41.0[illegible]	47.00	3.40	45.00	8.00	15.00	[illegible]
Central	Central	KS	194	87.00	1.00	5.00	46.00	0.00	10.00	40.00
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.00	51.00	0.01	26.00	41.00

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Cook	St. Louis	MO	335	0.00	100.00	0.00	45.00	0.00	15.00	00
Earl Nash	Noxubee	MS	509	1.00	98.00	1.00	51.00	0.20	6.50	100.00
Treatment school means			439	30.60	55.60	10.60	48.76	7.68	12.31	70.70
Control										
Newby[a]	Mooresville	IN	252	98.00	0.00	0.00	53.00	0.00	10.00	35.00
Jamestown	Guilford	NC	516	41.00	47.00	3.40	45.00	8.00	15.00	39.00
Central	Central	KS	194	87.00	1.00	5.00	46.00	0.00	10.00	40.00
Walnut Cove[a]	Walnut Cove	NC	355	79.00	17.00	1.00	51.00	0.01	26.00	41.00
Bluford	Guilford	NC	420	9.50	86.70	1.20	48.00	0.00	13.00	42.00
Greenwood	Bessemer	AL	375	13.00	75.00	10.00	52.00	11.00	13.00	76.00
Gulfview	Hancock	MS	590	95.00	3.00	1.00	49.00	0.00	6.70	80.00
Eutaw	Greene	AL	338	0.57	98.50	0.00	49.00	0.00	4.00	86.00
C. F. Hard	Bessemer	AL	619	0.20	99.70	0.00	46.00	0.00	6.00	90.00
Daniel Webster	Chicago	IL	671	0.00	100.00	0.00	44.00	0.00	5.00	95.00
Augustin Lara	Chicago	IL	616	0.50	0.40	98.60	50.00	54.00	8.70	95.70
Edward E. Dunne	Chicago	IL	531	0.00	100.00	0.00	48.00	0.00	3.80	96.00
Bunche	Chicago	IL	643	0.00	100.00	0.00	68.00	0.00	9.70	98.00
Cupples	St. Louis	MO	171	0.00	100.00	0.00	56.00	0.00	2.00	98.00
Dewey	Chicago	IL	616	0.00	98.00	0.00	65.00	25.00	6.30	98.00
Scullin	St. Louis	MO	282	0.00	100.00	0.00	42.00	0.00	13.00	98.00
M. E. Lewis[a]	Sparta	GA	631	1.16	97.67	0.00	50.60	0.16	17.30	99.00
Sigel	St. Louis	MO	340	8.10	85.30	1.80	45.00	18.00	16.00	100.00
South Delta	South Delta	MS	769	5.00	94.00	1.00	53.00	0.00	4.00	100.00
Stanfield	Stanfield	AZ	757	19.40	1.00	67.70	50.30	50.00	19.00	100.00
Control school means			484	22.90	65.20	9.50	50.55	8.31	10.43	80.34

Note. ESL = English as a second language.
[a]Selected in the pilot phase of the study.

TABLE 2
Comparison of Baseline Characteristics: Success for All K–2 Treatment Schools (N = 21) and Control Schools (N = 20)

Variable	Condition	*M*	*SD*	95% confidence interval for difference: Lower bound	Upper bound	*t*(40)
PPVT score	Control	89.25	8.40	−6.56	4.30	0.42
	Treatment	90.38	8.78			
Enrollment	Control	484.00	186	−78.79	169.48	0.46
	Treatment	439.00	206			
Female (%)	Control	50.54	6.47	−2.03	5.60	0.35
	Treatment	48.76	5.60			
Minority (%)	Control	77.10	35.78	−16.28	31.64	0.52
	Treatment	69.63	39.35			
ESL (%)	Control	8.31	16.49	−9.12	10.38	0.90
	Treatment	7.68	14.35			
Special education (%)	Control	10.43	6.14	−6.28	2.51	0.39
	Treatment	12.31	7.64			
Free lunch (%)	Control	80.34	25.14	−7.56	26.81	0.26
	Treatment	70.71	29.02			

Note. PPVT = Peabody Picture Vocabulary Test; ESL = English as a second language.

The sample is more African American and less Hispanic than Success for All schools nationally. Overall, 60.4% of the sample is African American (vs. about 40% of Success for All students), and 10.1% is Hispanic (vs. about 35% of Success for All schools). The percentage of White students (27%) is similar to that of Success for All schools (about 25%).

The results shown in Table 2 provide direct comparisons of the baseline characteristics of the K–2 treatment schools and the control schools. As the results indicate, the percentages of female, minority, special education, free-lunch, and English-as-a-second-language students were statistically equivalent across the treatment and control schools. Likewise, *t* tests focusing on school enrollment and baseline Peabody Picture Vocabulary Test (PPVT) outcomes for the schools showed no differences.

Therefore, the treatment and control samples were sufficiently well matched at baseline on key demographic characteristics and the PPVT pretest measure. Although the school sample includes a higher percentage of African American students and a smaller proportion of Latino students than other Success for All schools, it is similar with respect to percentage of free-lunch participants. The sample is also composed of schools from diverse locales, including high-poverty urban and rural schools across 11 states. In these respects, the sample selection process and randomization procedure appear to have produced a baseline sample of schools with good internal validity, in that there are no large, statistically significant treatment-control differences, and good external validity, in that the sample's demographic characteristics resemble those of the overall population of Success for All schools and a range of regional contexts are included, representing the national reach of the program.

Treatment Fidelity

Trainers from the Success for All Foundation made quarterly implementation visits to each school, as is customary in all implementations of the Success for All program. These visits established each school's fidelity to the Success for All model and provided trainers an opportunity to work with school staff in setting goals toward improving implementation. Many efforts were made to ensure the fidelity of the experimental treatment. As an ordinary part of the intervention, teachers in Success for All schools receive 3 days of training and then about 8 days of on-site follow-up during the first implementation year. Success for All Foundation trainers visit classrooms, meet with groups of teachers, look at data on children's progress, and provide feedback to school staff on implementation quality and outcomes. These procedures, followed in each

Success for All school, were used in the study schools in an attempt to obtain a high level of fidelity in regard to implementation.

As of January 2004, all K–2 classes in the schools were implementing their assigned treatments. There was some variability in implementation quality, which will be the subject of future analyses. For instance, several schools needed almost 1 year to understand and implement the program at a mechanical level, and others embraced the program immediately and are doing an excellent job. In a number of cases, the difficulties involved in recruiting schools and the last-minute recruitment that often took place significantly inhibited quality implementation, as many of the study schools (especially those in Chicago, St. Louis, and Guilford County, North Carolina) did not have time to complete the extensive planning typically engaged in by Success for All schools before the beginning of the school year.

In the control grades, teachers were repeatedly reminded to continue using their usual materials and approaches and not to use anything from Success for All. During the implementation visits, the trainers observed classrooms from the control grades. Specifically, these observations focused on whether the environment, instruction, and behaviors in the control classrooms resembled the characteristics of the Success for All classrooms. In no case did the trainers observe teachers in control classes implementing Success for All components. It is possible that some of the Success for All ideas or procedures did influence instruction in the control grades, but any such influence was apparently subtle. Instructional materials and core procedures were clearly distinct from each other in the treatment and control grades.

Measures

Students in kindergarten and first grade were pretested on the PPVT and then individually posttested on the Woodcock-Johnson measure by testers hired, trained, and supervised by the National Opinion Research Center at the University of Chicago. The 6 pilot schools were pretested in fall 2001 and posttested in spring 2002, and the 35 schools from the main sample were pretested in fall 2002 and posttested in spring 2003. The pilot and main samples were combined in the analyses.

Pretests

All children were individually assessed on the PPVT in fall 2001 (pilot sample) or fall 2002 (main sample) by testers hired, trained, and supervised by the National Opinion Research Center. The few children among whom Spanish was their dominant language were pretested in Spanish on the Test de Vocabulario en Imagenes Peabody.

Posttests

During the spring of 2002 (pilot sample) and spring of 2003 (main sample), students in the main longitudinal cohorts (beginning in kindergarten and first grade) were individually assessed on the four subtests of the Revised Woodcock Reading Mastery Tests (WMTR): Letter Identification, Word Identification, Word Attack, and Passage Comprehension. (Such testing was also scheduled to take place during each subsequent spring through 2005.) The WMTR was normed on a national sample of children, and the internal reliability coefficients for the four subtests used were .84, .97, .87, and .92, respectively. Children in the initial cohorts are being followed as long as they remain in the same school; retention does not change their cohort assignment. They are also being followed into special education. Children who entered Success for All or control schools after fall 2002 are to be posttested each year and included in analyses of cohort means. Children who are English-language learners but are taught in English will be posttested in English each year.

Because of the possibility that statistical outliers would exert an overly strong influence on the results, we screened each of the four dependent measures before conducting analyses. Applying Tukey's (1977) definition, we identified as potential statistical outliers any WMTR outcomes that were more than three interquartile ranges above the 75th percentile or below the 25th percentile. No values for the Word Identification, Word Attack, or Passage Comprehension outcomes met this criterion. However, inspection of the values for Letter Identification outcomes revealed a somewhat negatively skewed distribution in which 172 of the 177 outliers were more than three interquartile ranges below the 25th percentile and 2 outliers were more than three interquartile ranges above the 75th percentile. These 177 cases, though, did not appear to be erroneous and were equally distributed across the treatment ($n = 87$) and control ($n = 90$) conditions. Therefore, we did not delete the outlying cases or attempt to transform the values for the Letter Identification measure.

Results

In the earlier description of the sample, we concluded that the analysis of the baseline data showed few important differences between treatment and control schools and that the sample of schools was geographically diverse and generally representative of the population of Success for All schools. In discussing the results of our analyses of first-year achievement outcomes, we begin by assessing whether there was differential data and sample attrition between treatment and control schools or systematic attrition from the analytical sample that may have changed its characteristics relative to those of the baseline sample.

The final analytical sample was composed of 2,593 students in 21 K–2 Success for All treatment schools and 2,444 students in 20 control schools. Listwise deletion of cases involving missing student posttest data did not cause differential attrition rates by program condition, $\chi^2(1, N = 5{,}736) = 1.03, p = .31$; 87% of the baseline sample of 2,966 treatment students and 88% of the 2,770 baseline controls were included in the preliminary analyses. Data and sample attrition occurred for two reasons. Of the 699 students who were excluded from the analyses, 405 (58%) were dropped because they had moved out of the school before the posttests were administered and thus had no outcome data, and 294 (42%) remained in school but missed the spring posttest.

To further investigate the internal validity of the study, we compared the pretest scores of treatment and control students who were dropped from the analyses. No statistically significant difference was found between the treatment and the control students, $t(5{,}735) = -0.50, p = 0.62$ (two-tailed), suggesting that the initial academic abilities of the treatment and control group students who were dropped from our analyses were similar.

To address the issue of external validity, we compared students who were retained in the analyses with students who were not retained. Results showed that students who were retained had higher pretest scores than those who were not retained, $t(5, 735) = -4.29, p < .001$ (two-tailed). Also, not surprisingly, mobile students who had left the Success for All and control schools were overrepresented among those with missing data, $\chi^2(1, N = 5{,}736) = 3{,}151.70, p < .001$. Thus, both low-achieving and mobile students from the sample schools were underrepresented in the analyses, and this compromised the external validity of the study in two ways. First, because past quasi-experimental evidence has consistently shown that Success for All tends to have the most profound educational effects on students who are struggling academically (Slavin & Madden, 2001), omission of low-achieving students with missing posttest data who remained in the Success for All schools is most likely to result in downward biases of treatment effect estimates. Second, the fact that the primary missing data mechanism was mobility limits generalization to nonmobile students who remained in the baseline treatment and control schools.

While we concede these limitations, it is also the case that data attrition claimed a total of only 12% of the baseline sample and that mobile students represented only 7% of the overall sample. Furthermore, there was no conflict in this experiment between random assignment of treatment and data missing at random. That is, among the data observations, those involving treatment condition assignments had covariate distributions similar to those involving control condition assignments. As noted by Rubin (1976) and Little and Rubin (1987), the missing data process is "ignorable" if, conditional on treatment and fully observed covariates, the data are missing at random.

Hierarchical Linear Model Analyses of First-Year Treatment Effects

Cluster randomized trials (CRTs) in education generally randomize at the level of the school or classroom and collect data at the level of the student. In many respects, they represent the optimal design for school-based and classroom-based interventions. They address practical problems, including the potential difficulties of randomizing individual teachers within schools or students within classrooms to alternate treatments, and they are often well aligned with the theory of how educational interventions work best: as coordinated, systemic initiatives delivered via organizational-level elements acting in concert. Although, in education, greater attention has been paid to these designs in recent years (Boruch et al., 2004), methodological work related to proper specification of impact estimates from CRTs is still evolving.

Estimation of treatment effects at the level of the cluster that was randomized is the most appropriate method (Donner & Klar, 2000; Murray, 1998; Raudenbush, 1997). When the number of clusters is small, however, this strategy will not be efficient and will lack the necessary statistical power. If clustering is simply ignored and the analysis is conducted at the level of the individual student, this will create the illusion that statistical power has been substantially increased. However, these standard tests of statistical significance, which assume that the outcome for a particular student is completely unrelated to (or independent from) that for any other student, are inappropriate for CRTs. The reason is that, in CRTs, two students randomized together within any one classroom or school are more likely to respond in a similar manner than two students randomized from different clusters. If one computed the standard errors for a CRT as if individuals had been randomized, the outcome would understate the true standard errors substantially, thereby lending a false sense of confidence to the impact estimates. As Cornfield (1978, p. 101) noted: "Randomization by group accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception."

A relatively recently proposed analytical strategy for the analysis of CRTs is use of a hierarchical linear model (Raudenbush, 1997). In this formulation, one may simultaneously account for both student- and school-level sources of variability in outcomes by specifying a two-level hierarchical model that estimates the school-level effect of random assignment. Our fully specified Level 1, or within-school, model nested students within schools and involved an indicator of students' baseline grade level (–0.5 = kindergarten, 0.5 = first grade). The linear model for this level of the analysis is written as

$$Y_{ij} = \beta_{0j} + \beta_{1j}(\text{GRADE})_{ij} + r_{ij},$$

which represents the spring posttest achievement for student i in school j regressed on grade level plus the Level 1 residual variance, r_{ij}, that remains unexplained after student grade level has been taken into account.

In this model, each student's grade level is centered around zero, thus controlling for school-to-school differences in the proportions of kindergartners and first graders. With grade coded as –0.5 (kindergarten) or 0.5 (first grade), the Level 2 school-specific intercept represents the overall average school performance of kindergarten and first-grade students. We treat the within-school grade-level gap—the difference between the post-test scores of kindergarten and first-grade students in school j—as fixed at Level 2 because it is intended only as a covariate and we have no empirical or theoretical reason to model this source of between-school variability as an outcome.

At Level 2 of the model, we estimate the cluster-level impact of Success for All treatment assignment on the mean posttest achievement outcome in school j. As suggested by the work of Bloom, Bos, and Lee (1999) and Raudenbush (1997), we included a school-level covariate, school mean PPVT pretest score, to help reduce the unexplained variance in outcomes and to improve the power and precision of our treatment effect estimates.[2] The fully specified Level 2 model is written as

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{MEANPPVT})_j + \gamma_{02}(\text{SFA})_j + u_{0j},$$

$$\beta_{1j} = \gamma_{10},$$

where the mean posttest intercept for school j, β_{0j}, is regressed on the school-level mean PPVT score, the Success for All treatment indicator, and a residual, u_{0j}. The within-school posttest difference between kindergarten and first-grade students, β_{1j}, is specified as fixed, predicted only by an intercept.[3]

For each of the four achievement outcomes, we specified a series of four multilevel models. The preliminary unconditional model partitions the variation in outcomes among students and schools and is used as a basis for comparing the fit of subsequent models, which introduce student- and school-level predictors of posttest outcomes. Model 1 adds the student grade-level covariate as a predictor of the posttest. This model serves as a referent for the subsequent models when calculating the percentage of variance explained for the Level 2 school mean achievement intercepts. Model 2 introduces the school-level average PPVT pretest score as a predictor of achievement intercepts. Finally, Model 3 adds the Success for All treatment assignment indicator as a predictor of school mean achievement.[4]

Letter Identification Outcomes

The first series of multilevel models, shown in Table 3, assessed student and school-level effects on the WMTR Letter Identification posttest. The unconditional model with no student- or school-level predictors revealed the overall average value on the outcome measure, partitioned the variance in the outcome into its between- and within-school components, and indicated whether there was a statistically significant amount of between-school variance to model with independent variables. In terms of Letter Identification outcomes, the unconditional model yielded an average spring score of 0.11. The intraclass correlation coefficient was 0.08, which indicated that 8% of the variance in the Letter Identification posttest was between schools and that there was a statistically significant, $\chi^2(40, N = 41) = 498.65$, amount of Level 2 variability potentially explainable by school-level characteristics.

In the version of Model 1 shown in Table 3, we included the student-level grade indicator as a predictor; this explained 21% of the within-school posttest variability. The average within-school difference between kindergarten and first-grade students on standardized posttest outcomes was 0.86, or nearly one standard deviation. After controlling for the grade-level fixed effect, the school-specific intercept of 0.01, which represented the overall average school performance of kindergarten and first-grade students in school *j*, continued to show statistically significant random variation across schools. In the ensuing models, we attempted to explain this random variation across schools using the aggregate school-level PPVT pretest score and, most important, the Success for All treatment indicator.

Model 2 added school mean PPVT score as a predictor of mean achievement (see Table 3). The statistically significant coefficient for mean PPVT pretest score suggests that schools with higher such scores also tended to have higher Letter Identification posttest scores. Specifically, a one-standard-deviation increase on the PPVT pretest was associated with an increase of nearly one quarter of a standard deviation (0.22) on the posttest. By including school mean PPVT pretest score as a Level 2 predictor, the model explained 57% of the variability in school mean achievement.

Finally, Model 3 introduced the Success for All treatment indicator as a predictor of school mean achievement. This predictor showed no statistically significant first-year school-level effect of assignment to Success for All among kindergarten and or first-grade students. The Success for All coefficient for school mean achievement indicates that the effect of treatment assignment on Letter Identification posttest scores was essentially equal to a standardized mean difference, or effect size, of zero (−0.03).

Word Identification Outcomes

In Table 4, we model student- and school-level effects on WMTR Word Identification posttest scores. The preliminary unconditional model indicated that 10% of the variance in posttest scores was between schools and that there was a statistically significant, $\chi^2(40, N = 41) = 648.56$, amount of Level 2 variability in school mean achievement that could be explained by school-level predictors.

In Model 1 (see Table 4), we included the grade-level dummy code as a predictor. This independent variable explained 44% of the within-school posttest variance. The coefficient of 1.25 for grade level indicated that the average within-school difference in posttest scores between kindergarten and first-grade students was approximately 1.25 standard deviation units. After accounting for differences across schools in proportions of kindergarten and first-grade students, the model revealed a statistically significant amount of additional random variation across schools.

In Model 2, we explained this school-level variation using school-level average PPVT score as a predictor of school mean achievement. This measure was a statistically significant predictor of school mean achievement, explaining 39% of the between-school variability in outcomes.

In Model 3, we introduced the Success for All treatment indicator as a predictor of school mean achievement. In the case of Word Identification outcomes, there was no statistically significant first-year effect of school-level assignment to Success for All. The Success for All coefficient of 0.04 for mean achievement outcome suggested that the school-level effect size for assignment to treatment was close to zero.

TABLE 3
Multilevel Models Predicting Letter Identification Outcomes

	Unconditional model			Model 1			Model 2			Model 3		
Type of measure	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*
Fixed effect												
School mean achievement												
Intercept	0.11	0.05	0.24	0.01	0.05	0.32	0.02	0.03	0.50	0.02	0.03	0.50
Mean PPVT pretest							0.22**	0.03	6.62	0.22**	0.03	6.55
SFA assignment										−0.03	0.06	−0.48
Grade												
Intercept				0.86**	0.04	19.65	0.86**	0.04	19.69	0.86**	0.04	19.69
Random effect	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*
School mean achievement	0.08	498.65	40	0.08	595.09	40	0.03	285.33	39	0.04	282.64	38
Within-school variation	0.87			0.68			0.68			0.68		
Variance explained (%)												
Within-school variation				21			21			21		
School mean achievement							57			56		

Note. PPVT = Peabody Picture Vocabulary Test; SFA = Success for All.
**$p < .001$.

TABLE 4
Multilevel Models Predicting Word Identification Outcomes

Type of measure	Unconditional model			Model 1			Model 2			Model 3		
	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*
Fixed effect												
School mean achievement												
Intercept	−0.02	0.05	−0.42	−0.02	0.05	−0.35	−0.02	0.04	−0.41	−0.02	0.04	−0.41
Mean PPVT pretest							0.21**	0.04	5.29	0.20**	0.04	5.23
SFA assignment										0.04	0.08	0.47
Grade												
Intercept				1.25**	0.04	29.65	1.25**	0.04	29.70	1.25**	0.04	29.70
Random effect	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*
School mean achievement	0.10	648.56	40	0.10	1,119.38	40	0.06	714.53	39	0.07	710.67	38
Within-school variation	0.88			0.49			0.49			0.49		
Variance explained (%)												
Within-school variation				44			44			44		
School mean achievement							39			37		

Note. PPVT = Peabody Picture Vocabulary Test; SFA = Success for All.
**$p < .001$.

Word Attack Outcomes

In Table 5, we present the results from the hierarchical models predicting WMTR Word Attack outcomes. Again, we begin with the unconditional model, which indicates that the average posttest Word Attack score was –0.34. The intraclass correlation coefficient for this model was 0.12, indicating that 12% of the variance in the Word Attack posttest was between schools. The estimate for the school mean achievement random effect indicated a statistically significant, $\chi^2(40, N = 41) = 772.37$, amount of Level 2 variability between schools.

Model 1 included the student grade-level indicator as a predictor. This predictor explained 25% of the within-school posttest variance. In Model 2, we used school-level average PPVT score to explain the additional variation in school mean achievement that existed after controlling for differences across schools in their proportions of kindergarten and first-grade students. Aggregate PPVT pretest score was a statistically significant predictor of school mean achievement. The positive relationship between school mean PPVT score and posttest score suggested that a one-standard-deviation increase on the pretest was related to an increase of nearly one quarter of a standard deviation on the posttest. School mean PPVT pretest score explained 37% of the between-school variability in school mean achievement.

Finally, Model 3 introduced the Success for All treatment indicator as a predictor of school mean achievement. In the case of Word Attack outcomes, this model revealed a statistically significant first-year effect of school-level assignment to Success for All. The Success for All coefficient of 0.22 for mean achievement outcome indicated that the magnitude of the school-level effect of assignment to Success for All was nearly one quarter of a standard deviation. By adding the Success for All treatment indicator, we explained an additional 10% of the variability across schools in school mean achievement.

Passage Comprehension Outcomes

Results for the final outcome measure, Passage Comprehension, are shown in Table 6. The unconditional model indicated that the average spring Passage Comprehension score was –0.01. Similar to the other models, 10% of the variation in outcomes was between schools. This represented a statistically significant amount of between-school variability for the school mean achievement intercepts, $\chi^2(40, N = 41) = 578.70$.

In Model 1, we added the student grade-level indicator as a predictor; this predictor accounted for 39% of the within-school variation on the posttest. In Model 2, we attempted to explain the random variation across schools in their mean achievement intercepts using school-level average PPVT score as a predictor. The schools' average PPVT pretest scores explained 48% of the between-school variability in mean posttest outcomes.

Finally, Model 3 included the Success for All treatment indicator as a predictor of school mean achievement. In regard to Passage Comprehension outcomes, this model showed no statistically significant first-year effects of school-level assignment to Success for All. The coefficient of –0.03 for the Success for All dummy code indicated that the effect size for assignment to the Success for All treatment was essentially zero.[5]

Discussion

The first-year outcomes of the Success for All national randomized trial are noteworthy for several reasons. First, the selection and randomization processes worked well. With considerable effort and expense, we were able to obtain the cooperation of a sufficient number of Success for All and control schools to provide an acceptable level of statistical power to detect school-level effects within a multilevel-model framework. No matter how carefully drawn, a sample of 41 schools is not likely to represent the population of more than 1,400 Success of All schools with great precision. However, the process does seem to have developed a sample of schools that is similar to the overall population of Success for All schools with respect to poverty level and that is geographically diverse, spread across 11 states. Randomization produced control and K–2 treatment samples that were reasonably well matched on a variety of baseline characteristics, including demographics and PPVT pretest scores. No statistically significant baseline differences were detected at the school level.

Second, the data and sample attrition over the first year of the study had minimal impact on the good external and internal validity achieved through the sample selection and randomization procedures. There was no differential rate of data attrition or mobility between Success for All K–2

TABLE 5

Multilevel Models Predicting Word Attack Outcomes

Type of measure	Unconditional model			Model 1			Model 2			Model 3		
	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*
Fixed effect												
School mean achievement												
Intercept	−0.34	0.06	−0.62	−0.03	0.05	0.58	−0.03	0.04	−0.68	−0.03	0.04	−0.74
Mean PPVT pretest							0.22**	0.05	4.86	0.21**	0.04	4.67
SFA assignment										0.22*	0.08	2.72
Grade												
Intercept				0.94**	0.06	17.15	0.94**	0.06	17.17	0.94**	0.05	17.18
Random effect	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*
School mean achievement	0.12	772.37	40	0.12	1,008.07	40	0.08	654.22	39	0.06	531.03	38
Within-school variation	0.87			0.65			0.65			0.65		
Variance explained (%)												
Within-school variation				25			25			25		
School mean achievement							37			47		

Note. PPVT = Peabody Picture Vocabulary Test; SFA = Success for All.
$^*p < .05$; $^{**}p < .001$.

TABLE 6

Multilevel Models Predicting Passage Comprehension Outcomes

Type of measure	Unconditional model			Model 1			Model 2			Model 3		
	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*
Fixed effect												
School mean achievement												
Intercept	−0.01	0.05	−0.29	−0.01	0.05	−0.22	−0.01	0.04	−0.24	−0.01	0.04	−0.24
Mean PPVT pretest							0.22**	0.04	5.56	0.22**	0.04	5.61
SFA assignment										−0.03	0.07	−0.50
Grade												
Intercept				1.18**	0.05	25.23	1.18**	0.05	25.27	1.18**	0.05	25.26
Random effect	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*
School mean achievement	0.10	578.70	40	0.09	902.21	40	0.05	484.16	39	0.05	481.96	38
Within-school variation	0.90			0.55			0.55			0.55		
Variance explained (%)												
Within-school variation				39			39			39		
School mean achievement							48			47		

Note. PPVT = Peabody Picture Vocabulary Test; SFA = Success for All.

**$p < .001$.

treatment schools and control schools, and comparisons of students omitted from the analyses revealed no treatment-control differences with respect to pretest scores. However, statistically significant differences were revealed in comparisons of pretest scores among students who were omitted from the analyses—across both Success for All and control schools—and students who remained in the analytical sample. Specifically, mobile students and students with lower PPVT pretest scores were overrepresented among those dropped from the analytical sample. The data attrition rate, although quite low, did somewhat compromise the external validity of the findings. In future rounds of the study, as data attrition may become more severe, we will use strategies, including tracking and testing mobile students who move into Success for All and control schools, to retain samples that are representative of the student populations at the study schools.

Third, treatment fidelity and Success for All implementation quality seemed to be reasonably good. In some schools, the tight deadlines involved in the selection and randomization process appear to have allowed insufficient time for the program to become established and flourish. These qualitative differences in implementation quality will be an important subject of future work. Our ongoing measurement of the implementation of school and classroom practices in Success for All and control schools will allow us not only to describe implementation variability across schools but also to estimate (through application of statistical models that estimate complier average causal effects) the causal effects of compliance with the Success for All components on achievement outcomes.

Finally, the pattern of first-year treatment effects we found appears to be consistent with previous quasi-experimental work on Success for All, the Success for All program theory, and more general research and theory on the development of young children's emergent literacy skills. We found effects of both statistical and practical significance on the Word Attack posttest but did not find such effects on Letter Identification, Word Identification, or Passage Comprehension. The magnitude of the school-level effect of assignment to the Success for All intervention was equal to an effect size (*d*) of 0.22, or 2.11 months of additional learning relative to control schools.[6] Although previous quasi-experimental studies of Success for All have shown effects of greater magnitudes, the same pattern has existed. That is, the strongest first-year effects of the program tend to be in the domain tested by the Word Attack measure. In later years, the phonetic and structural analysis skills tapped by the Word Attack test help Success for All children develop more advanced comprehension and reading skills, especially those measured by the Passage Comprehension test. The treatment effects in these other skill areas typically have become more pronounced after children's exposure to 2 or more years of the Success for All program.

This pattern of effects fits the theory of Success for All, which focuses on the development of "reading roots" in the early grades. A strong focus of the reading roots component of Success for All is to increase children's ability to hear sounds within words (phonemic awareness) and to use phonetic strategies to decode text. Although establishing a love of reading and nurturing a child's literacy development within the context of meaningful literature are key components of developing these early skills, story-related activities and direct instruction in reading comprehension are more clearly stressed in the later grades within the Success for All "reading wings" component. This program theory for developing children's literacy skills is consistent with more general theories of how young children develop as emergent readers (Snow, Burns, & Griffin, 1998). Specifically, powerful decoding strategies and phonemic awareness, as stressed by the kindergarten and first-grade Success for All program, are key building blocks on which children can develop a broader range of skills.

Implications and the Future of the National Randomized Field Trial

In future research, we will examine how well the program theory matches the outcomes observed in the various reading assessments. In addition, we will assess the extent to which schoolwide Success for All programs affect other important and more general school outcomes, including special education referrals and attendance and retention rates. The first-year investigation described here establishes a strong foundation for this future work.

The study also represented a response to the many doubts that have been raised about the viability and appropriateness of randomized experiments in school settings (Cook & Payne, 2002).

Because the study was a randomized field trial rather than a relatively artificial laboratory experiment, the results have strong external validity and relevance for policy and practice. Furthermore, combined with future survey and qualitative data regarding implementation and process, this research will open the experimental "black box" and generate additional information that will be helpful to policymakers, practitioners, and scholars alike. Of course, we will continue to generate estimates of treatment effects in subsequent analyses, but additional information will help inform questions regarding how the effects were, or were not, attained. These are the types of data that can, ultimately, improve programs and practices. Through this information and the careful description of the study design, we also hope to continue to provide insights that will help researchers design and implement future randomized field trials focusing on educational interventions, especially those that involve place-based cluster random assignment of schools.

As we have noted, the process of randomization was not easy or inexpensive, but a prominent goal of this study was to recruit a group of schools agreeable to the idea of randomization. In this way, experimentation was viewed as a partnership with school personnel rather than as a process imposed on practitioners. The population of schools to which we wish to generalize comprises those that would like to adopt and implement Success for All. To foist reform on schools unwilling to implement it would not be consistent with the Success for All model or with how reform generally occurs in education, and such a study would have limited generalizability. Rather, this project tied together two central themes of educational research and policy today: the scaling up, or replication, of school-based interventions and the development of high-quality evidence of their causal effects. The first-year outcomes described here establish that cluster randomized field trials involving nationally replicated school-based interventions are both possible and desirable for producing unbiased estimates of the effects of educational treatments.

Appendix: Major Elements of Success for All

Success for All is a schoolwide program for students in prekindergarten to Grade 6 that organizes resources in an attempt to ensure that all children will be successful in reading from the beginning of their time in school and will never begin the process of falling behind. The emphasis of the program is on prevention and early, intensive intervention designed to detect and resolve reading problems as early as possible, before they become serious. The main elements of the program are as follows.

Schoolwide curriculum: During reading periods, students are regrouped across age lines so that each reading class contains students all at one reading level. Use of tutors as reading teachers during reading time reduces the size of most reading classes to about 20. The K–1 reading program emphasizes language and comprehension skills, phonics, sound blending, and use of shared stories that students read to one another in pairs. The shared stories combine teacher-read material with phonetically regular student material to teach decoding and comprehension in the context of meaningful, engaging stories. In Grades 2–6, students use novels or basals but not workbooks. This program emphasizes cooperative learning activities built around partner reading; identification of characters, settings, problems, and problem solutions in narratives; story summarization; writing; and direct instruction in reading comprehension skills. At all levels, students are required to read books of their own choice for 20 minutes at home each evening. Classroom libraries of trade books are provided for this purpose. Cooperative learning programs in writing/language arts are used in Grades K–6.

Tutors: In Grades 1–3, specially trained certified teachers and paraprofessionals work one to one with any students who are failing to keep up with their classmates in reading. Tutorial instruction is closely coordinated with regular classroom instruction. It takes place 20 minutes daily during times other than reading periods.

Preschool and kindergarten: The preschool and kindergarten programs in Success for All emphasize language development, readiness, and self-concept. Preschools and kindergartens use thematic units, language development activities, and a program called Story Telling and Retelling.

Quarterly assessments: Students in Grades 1–6 are assessed every quarter to determine whether they are making adequate progress in reading. This information is used to suggest alternate teaching strategies in the regular classroom, changes in reading group placement, provision of tutoring services, or other means of meeting students' needs.

Family support team: A family support team works in each school to help support parents in ensuring the success of their children, focusing on parent education, parent involvement, attendance, and student behavior. This team is composed of existing or additional staff such as parent liaisons, social workers, counselors, and vice principals.

Facilitator: A program facilitator works with teachers to help them implement the reading program, manages the 8-week assessments, assists the family support team, ensures that all staff are communicating with each other, and helps the staff as a whole make certain that every child is showing adequate progress.

Training: Success for All provides extensive training to help all teachers use the program effectively. New schools begin with weeklong training for the principal and facilitator, followed by a 3-day workshop before school opening for all staff. Implementation visits and additional training sessions are then provided throughout the first year and continue on a gradually diminishing basis through the second, third, and subsequent years. In addition, school facilitators provide training and follow-up daily to all staff.

Notes

This research was supported by grants from the Institute of Education Sciences (IES), U.S. Department of Education (R306 S000009 and R-11 7-40005). However, the opinions expressed are those of the authors and do not necessarily represent IES positions or policies.

[1]The oversight committee members are Ronald Ferguson, Harvard University; Steve Raudenbush, University of Michigan; Rebecca Maynard, University of Pennsylvania; Jonathan Crane, Progressive Policy Institute; and Kent McGuire, Temple University and Manpower Demonstration Research Corporation.

[2]We formulated other multilevel models that included the broader array of school-level covariates listed in Table 2. After inclusion of the school mean pretest covariate, though, these more complex models did not explain appreciably more between-school variance and did not improve the precision of the Success for All treatment effect estimates. For these reasons, we used the more parsimonious models presented.

[3]The statistical precision of the design can be expressed in terms of a minimum detectable effect, or the smallest treatment effect that can be detected with confidence. As noted by Bloom (in press), this parameter, which is a multiple of the impact estimator's standard error, depends on whether a one- or two-tailed test of statistical significance is used; the alpha level of statistical significance to which the result of the significance test will be compared; the desired statistical power, $1 - \beta$; and the number of degrees of freedom of the test, which equals the number of clusters, J, minus 2 (assuming a two-group experimental design and no covariates).

The minimum detectable effect for our design is calculated for a two-tailed t test with an alpha level of $p < .05$; power, $1 - \beta$ equal to 0.80; and degrees of freedom equal to $J = 41$ schools minus 3 (a two-group experimental design with the school mean PPVT pretest covariate). Referring to the data shown in Tables 3 through 6 regarding the Success for All impact estimators' standard errors, which ranged from 0.06 to 0.08, and employing Bloom's (in press) minimum detectable effect multiplier, we calculated minimum detectable effects (d values) of approximately 0.17 to 0.22. That is, our design had adequate power to detect school-level treatment-control differences of at least 0.17 to 0.22 standard deviations.

[4]All multilevel models were estimated via the HLM software's restricted maximum likelihood estimation procedure (Raudenbush, Bryk, Cheong, & Congdon, 2000).

[5]We conducted analyses of empirical Bayes residuals for each of the four outcomes assessed in the hierarchical linear models. As a result of the inclusion of potential outliers in the analyses of Letter Identification outcomes, we were particularly concerned about checking for normality at Level 2 of the analytical models. We computed Mahalanobis distance measures for each of the 41 schools included in the analyses. This measure assesses the distance between the residual estimates for each group relative to the expected distance based on the specified model. The measure, which has a chi-square distribution with $Q + 1$ degrees of freedom when data are normal, provides a summary of the degree of departure of random effects from normality and permits identification of outliers. After controlling the Level 1 grade-level indicator, we calculated the Mahalanobis distance measures and plotted them against the expected values. Mahalanobis distance plots for the hierarchical linear models of the four outcomes revealed no unusual departures from normality.

[6]We estimated additional months of learning through a two-step process. First, we referred to the coefficient for grade level of 0.94 from the hierarchical linear model of Word Attack outcomes (see Table 3) in an attempt to understand the change in outcomes associated with a one-grade-level difference. This one-grade-level difference on the Word Attack measure approximates the amount of growth that occurs across a single school

year, from the end of kindergarten through the end of first grade. Dividing the coefficient for the Success for All treatment effect of 0.22 (see Table 3) by this grade-level coefficient, we calculated the percentage of one grade level represented by the treatment effect (0.22/0.94 = 0.23). Second, we estimated the duration of 1 school year as 9 months. Multiplying the figure of 0.23 derived from the first step by 9, we converted the treatment effect coefficient into an estimate of additional months of learning (0.23 × 9 = 2.11).

References

Bloom, H. S. (in press). *Learning more from social experiments: Evolving analytic approaches.* New York: Russell Sage Foundation.

Bloom, H. S., Bos, J. M., & Lee, S.-W. (1999). Using cluster random assignment to measure program impacts: Statistical implications for the evaluation of education programs. *Evaluation Review, 23,* 445–469.

Bloom, H. S., Michaelopoulous, C., Hill, C. J., & Lei, Y. (2002). *Can non-experimental comparison group methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs?* New York: Manpower Demonstration Research Corporation.

Borman, G. D. (2003). Experiments for educational evaluation and improvement. *Peabody Journal of Education, 77*(4), 7–27.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73,* 125–230.

Boruch, R., de Moya, D., & Snyder, B. (2002). The importance of randomized field trials in education and related areas. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 50–79). Washington, DC: Brookings Institution.

Boruch, R., May, H., Turner, H., Lavenberg, J, Petrosino, A., de Moya, D., Grimshaw, J., & Foley, E. (2004). Estimating the effects of interventions that are deployed in many places: Place-randomized trials. *American Behavioral Scientist, 47,* 608–633.

Cook, T. D. (2002). Randomized experiments in education: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24,* 175–199.

Cook, T. D., & Payne, M. R. (2002). Objecting to the objections to using random assignment in educational research. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 150–178). Washington, DC: Brookings Institution.

Cornfield, J. (1978). Randomization by group: A formal analysis. *American Journal of Epidemiology, 108,* 100–102.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., Walker, D. F., & Weiner, S. S. (1980). *Toward reform of program evaluation: Aims, methods, and institutional arrangements.* San Francisco: Jossey-Bass.

Donner, A., & Klar, N. (2000). *Design and analysis of group randomization trials in health research.* London: Arnold.

Finn, J. D., & Achilles, C. M. (1999). Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis, 21,* 97–109.

Glazerman, S., Levy, D., & Myers, D. (2002). *Non-experimental replications of social experiments: A systematic review.* Washington, DC: Mathematica Policy Research.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181–1209.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Murray, D. M. (1998). *Design and analysis of group-randomized trials.* New York: Oxford University Press.

No Child Left Behind Act, Public Law No. 107-110, 115 Stat 1425 (2002).

Olson, L., & Viadero, D. (2002). Law mandates scientific base for research. *Education Week, 21*(20), pp. 1, 14–15.

Petrosino, A., Turpin-Petrosino, C., & Buehler, J. (2003). Scared Straight and other juvenile awareness programs for preventing juvenile delinquency: A systematic review of the randomized experimental evidence. *Annals of the American Academy of Political and Social Science, 589,* 41–62.

Pogrow, S. (2000). Success for All does not produce success for students. *Phi Delta Kappan, 82,* 67–81.

Ramey, C. T., & Campbell, F. A. (1984). Preventive education for high-risk children: Cognitive consequences of the Carolina Abecedarian Project. *American Journal of Mental Deficiency, 88,* 515–523.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2,* 173–185.

Raudenbush, S. W., Bryk, A. S., Cheong, Y., & Congdon, R. T. (2000). *HLM5: Hierarchical linear and nonlinear modeling.* Chicago: Scientific Software International.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592.

Schweinhart, L. J., Barnes, H. V., & Weikart, D. P. (1993). *Significant benefits: The High/Scope Perry*

Preschool Study through age 27. Ypsilanti, MI: High/Scope Press.

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

Slavin, R., & Madden, N. (2001). *One million children: Success for All.* Thousand Oaks, CA: Corwin Press.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children.* Washington, DC: Committee on the Prevention of Reading Difficulties in Young Children, National Research Council.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Walberg, H., & Greenberg, R. (1999). The Diogenes factor. *Phi Delta Kappan, 81,* 127–128.

Authors

GEOFFREY D. BORMAN is an Associate Professor, Departments of Educational Leadership and Policy Analysis, Educational Psychology, and Educational Policy Studies, University of Wisconsin, 1161D Educational Sciences Building, 1025 West Johnson Street, Madison, WI 53706; gborman@education.wisc.edu. His areas of specialization include policy, stratification, research methodology, and evaluation.

ROBERT E. SLAVIN is the Director of the Center for Data-Driven Reform in Education at Johns Hopkins University and Chairman of the Success for All Foundation, 200 West Towsontown Boulevard, Baltimore, MD 21204; rslavin@successforall.net. His areas of specialization include cooperative learning, comprehensive school reform, evidence-based policy, English language learners, and research methods.

ALAN CHEUNG is a Research Scientist at the Success for All Foundation, 200 West Towsontown Boulevard, Baltimore, MD 21204; acheung@successforall.net. His areas of specialization include research reviews and evaluation.

ANNE M. CHAMBERLAIN is a Research Scientist at the Success for All Foundation, 200 West Towsontown Boulevard, Baltimore, MD 21204; achamberlain@successforall.net. Her areas of specialization are program evaluation, mixed methodology, and participatory evaluation.

NANCY A. MADDEN is the President and CEO of the Success for All Foundation, 200 West Towsontown Boulevard, Baltimore, MD 21204; nmadden@successforall.net. Her areas of specialization include cooperative learning, comprehensive school reform, evidence-based policy, English-language learners, and research methods.

BETTE CHAMBERS is the Vice President of Development at the Success for All Foundation, 200 West Towsontown Boulevard, Baltimore, MD 21204; bchambers@successforall.net. Her areas of specialization are early childhood education, early reading, and the use of technology in education.

Manuscript received April 6, 2004
Revision received October 15, 2004
Accepted December 1, 2004

Evidence-Based Reform in Education: Promise and Pitfalls

Robert E. Slavin
Johns Hopkins University

Portions of this paper are adapted from Slavin, R.E. (2003), Evidence-based policies: Transforming educational practice and research. *Educational Researcher, 31* (7), 15-21. This paper was written under funding from the U.S. Department of Education (Grant No. OERI-R-117-D40005). However, any opinions expressed are those of the author and do not necessarily represent Department of Education positions or policies.

<LEVEL 1> Abstract

In this keynote address presented at the Mid-western Educational Research Association Annual Meeting in October, 2004, the author discusses the increasing interest of federal policy-makers on scientifically based research. A comparison between education and other disciplines is offered, and a proposal for increased rigor in educational research is offered.

Education is on the brink of a scientific revolution that has the potential to profoundly transform policy, practice, and research. Consider the following:

- In 1998, Congress appropriated $150 million per year to provide schools funds to adopt "proven, comprehensive reform models." This unprecedented legislation, introduced by Congressmen David Obey and John Porter, defined "proven" in terms of experimental-control comparisons on standards-based measures. To my knowledge, this was the first time in history that education funding anywhere has been linked directly to evidence of effectiveness (see Slavin, 1997). Comprehensive School Reform (CSR) funding progressively increased to $310 million annually, and has provided funding to more than 3000 mostly high-poverty schools.
- The Bush administration's main education initiative, No Child Left Behind, took the idea of scientifically based practice to an even higher level. The No Child Left Behind legislation refers to "scientifically-based research" 110 times. It defines "scientifically-based research" as "rigorous, systematic and objective

procedures to obtain valid knowledge," which includes research that "is evaluated using experimental or quasi-experimental designs…," preferably with random assignment. "Scientifically-based research" is intended to serve as the basis for a wide array of federally funded programs, especially Reading First programs for reading in grades K-3.

- Grover Whitehurst, the current director of the Institute for Education Research (IES) in the U.S. Department of Education, has taken a strong line in support of randomized experiments (Whitehurst, 2002). The U.S. Department of Education strategic plan for 2002-2007 anticipates having 75% of all OERI-funded research that addresses causal questions use random assignment designs by 2004 (previously, such research was less than 5% of causal research funded by The U.S. Department of Education). As a direct result, Congress significantly increased funding for education research. Research involving random assignment is now under way on early childhood programs, elementary and secondary reading, math, programs for English language learners, teacher professional development, after school remedial programs, and much more.

It is important to note that none of these policy developments have yet produced the revolution I am anticipating. These initiatives are too new to have had any impact on practice. Yet these and other developments, if not yet proven, still create the potential for changes with far-reaching consequences. It is possible that these policy reforms could set in motion a process of research and development on programs and practices affecting children everywhere. This process could create the kind of progressive, systematic improvement over time that has characterized successful parts of our economy and society throughout the 20th century, in fields such as medicine, agriculture, transportation, and technology. In each of these fields, processes of development, rigorous evaluation, and dissemination have produced a pace of innovation and improvement that is unprecedented in history (see Shavelson & Towne, 2002). These innovations have transformed the world. Yet education has failed to embrace this dynamic, and as a result, education moves from fad to fad. Educational practice does change over time, but the change process more resembles the pendulum swings of taste characteristic of art or fashion (think hemlines) rather than the progressive improvements characteristic of science and technology (see Slavin, 1989).

<LEVEL 1> Welcome to the 20th Century

At the dawn of the 21st century, education is finally being dragged, kicking and screaming, into the 20th century. The scientific revolution that utterly transformed medicine, agriculture, transportation, technology, and other fields early in the 20th century almost completely bypassed the field of education. If Rip Van Winkle had been a physician, a farmer, or an engineer in the 19th century, gone to sleep, and awoke today, he would be unemployable. If he had been a good primary school teacher in the nineteenth century, he'd probably be a good primary school teacher today. It's not that we haven't learned anything since Rip Van Winkle's time. It's that applications of the findings of educational research remain haphazard, and that evidence is respected only occasionally, and only if it happens to correspond to current educational or political fashions.

Early in the 20th century, the practice of medicine was at a similar point. For example, research had long since identified the importance of bacteria in disease, and by 1865 Joseph Lister had demonstrated the effectiveness of antiseptic procedures in surgery. In the 1890's, William Halsted at Johns Hopkins University introduced rubber gloves, gauze masks, and steam sterilization of surgical instruments, and demonstrated the effectiveness of these procedures. Yet it took thirty years to convince tradition-bound physicians to use sterile procedures. If he dropped his scalpel, a physician in 1910 was as likely as not to give it a quick wipe and carry on.

Today, of course, the linkage between research and practice in medicine is so tight that no physician would dream of ignoring the findings of rigorous research. Because medical practice is so closely based on medical research, funding for medical research is vast, and advances in medicine take place at breathtaking speed. My father's cardiologist recommended that he wait a few years to have a necessary heart valve operation because he was sure that within that short span of time, research would advance far enough to make the wait worthwhile. As it turned out, he was right.

The most important reason for the extraordinary advances in medicine, agriculture, and other fields is the acceptance by practitioners of evidence as the basis for practice. In particular, it is the randomized clinical trial, more than any single medical breakthrough, that has transformed medicine (Doll, 1998). In a randomized clinical trial, patients are assigned at random to receive one treatment or another, such as a drug or a placebo. Because of random assignment, it can be assumed with an adequate number of subjects that any differences seen in outcomes are due to the treatment, not to any extraneous factors. Replicated experiments of this kind can establish beyond any reasonable doubt the effectiveness (or lack thereof) of treatments intended for applied use (see Boruch, 1997).

<LEVEL 3> Experiments in Education

In education, experiments are not uncommon, but they are usually brief, artificial experiments on topics of theoretical more than practical interest, often involving hapless college sophomores. Far more rare are experiments evaluating treatments of practical interest studied over a full school year or more. I write an educational psychology textbook (Slavin, 2003) that is full of research findings of all kinds, findings that are valuable in advancing theory and potentially valuable to teachers in understanding their craft. Yet the brief experiments, correlational studies, and descriptive studies that yield most of the information presented in my text or any other educational psychology text do not collectively add up to school reform. They are suggestions about how to think about daily teaching problems, not guides to the larger questions educators and policymakers must answer. Imagine that research in cardiology described heart function and carried out small scale laboratory studies, but never developed and tested an artificial heart valve. If this were the case, I'd be an orphan. Imagine that agricultural research studied plant growth and diseases, but never developed and tested new disease-resistant crops. Educational research has produced many rigorous and meaningful studies of basic principles of practice, but very few rigorous studies of programs and practices that could serve as a solid base for policy and practice, and has had little respect for the studies of this kind that do exist. Because of this, policy makers have rarely seen the relevance of research to the decisions they have to make, and therefore have provided minimal funding for research. This has led to a declining spiral, as inadequate investments in research lead to a dearth of the kind of large-scale, definitive research that policy makers would feel to be valuable, making these policy makers unwilling to invest in large-scale, definitive research.

<LEVEL 3> Shifting Policy Perspectives

The dramatic changes in federal education policies I mentioned earlier could potentially reverse this declining spiral. If the new funding flowing into research in the U.S. can produce some notable successes, we could have an ascending spiral: rigorous research demonstrating positive effects of replicable programs on important student outcomes would lead to increasing funding for such research which would lead to more and better research and therefore more funding. More importantly, millions of children would benefit in the fairly near term. Once we establish replicable paradigms for development, rigorous evaluation, replication, and dissemination, these mechanisms could be applied to any educational intervention or policy. Imagine that there were programs under way all the time to develop, evaluate, and disseminate new programs in every subject and every grade level, as well as programs on school-to-work transitions, special education, gifted programs, dropout prevention, programs for English language learners, race relations programs, drug abuse prevention, violence prevention, and so on. Every one of these areas lends itself to a development-evaluation-dissemination paradigm, as would many more. Over time, each area would experience the step-by-step, irreversible progress characteristic of medicine and agriculture, because innovations would be held to strict standards of evaluation before being recommended for wide scale use.

<LEVEL 3> Research Designs

The scientific revolution in education will only take hold and produce its desired impacts if research in fact begins to focus on replicable programs and practices central to education policy and teaching, and if it in fact employs research methods that meet the highest standards of rigor.

This begs an important question: what kinds of research are necessary to produce findings of sufficient rigor to justify faith in the meaning of their outcomes?

Of course, all sorts of research designs are appropriate for various purposes, from description to theory building to hypothesis testing. However, leaders in the current administration and many other educational researchers throughout the world (see Angrist, 2004) have been arguing that nothing less than *randomized* experiments will do for evaluations of educational interventions and policies. When we want to know the outcome of choosing program X instead of program Y, there is no equivalent substitute for a randomized experiment.

<LEVEL 4> Randomized experiments

The difference in the value of randomized and well-matched experiments relates primarily to the problem of selection bias. In a matched experiment, it is always possible that observed differences are due not to treatments, but to the fact that one set of schools or teachers was willing to implement a given treatment while another was not, or that a given set of students selected themselves or were selected into a given treatment while others were not.

When selection bias is a possibility at the student level, there are few if any alternatives to random assignment, because unmeasured (often, unmeasurable) pre-existing differences are highly likely to be alternative explanations for study findings. For example, consider studies of after school or summer school programs. If a researcher simply compared students attending such programs to those not attending who were similar in pretest scores or demographic factors, it is very likely that unmeasured factors such as student motivation, parents' support for education, or other consequential factors could explain any gains observed, because the more motivated children are more likely to show up. Similarly, studies comparing children assigned to gifted or special education programs to students with similar pretest scores are likely to miss key selection factors that were known to whoever assigned the students but not measured. If one child with an IQ of 130 is assigned to a gifted program and another with the same IQ is not, it is likely that the children differ in motivation, conscientiousness, or other factors. In these kinds of situations, use of random assignment from within a selected pool is essential.

In contrast, there are situations in which it is teachers or schools that elect to implement a given treatment, but there is no selection bias that relates to the children. For example, a researcher might want to compare the achievement gains of children in classes using cooperative learning, or schools using comprehensive reform models, to the gains made by control groups. In such cases, random assignment of willing teachers or schools is still far preferable to matching, as matching leaves open the possibility that volunteer teachers or staffs are better than non-volunteers. However, the likely bias is much less than in the case of student self-selection. Aggregate pretest scores in an entire school, for example, should indicate how effective the current staff has been up to the present, so controlling for pretests in matched studies of existing schools or classes would control out much of the potential impact of having more willing teachers. For external validity, it is crucial to note that the findings of a well-matched experiment comparing volunteers to non-volunteers apply only to schools or teachers who volunteer, but the potential for bias is moderate (after controlling for pretests and demographic factors).

The importance of this discussion lies in the fact that randomized experiments of interventions applying to entire classrooms can be extremely difficult and expensive to do, and are sometimes impossible. My colleagues and I at Johns Hopkins University are doing a randomized evaluation of Success for All, a comprehensive reform model. Recruiting schools for this study was extremely difficult, even though we are offering substantial financial incentives to schools willing to be assigned at random to experimental or control groups. For the cost of doing this randomized study, we (and others) could have done two or three equally large-scale matched studies. It is at least arguable that replicated matched studies, done by different investigators in different places, might produce more valid and meaningful results than one definitive, once-in-a-lifetime randomized study.

Still, fully recognizing the difficulties of randomized experiments, I think they are nevertheless possible in most areas of policy-relevant program evaluation, and whenever they are possible, they should be used. Reviews of research in other fields have found that matched studies generally find stronger outcomes than randomized studies, although usually in the same direction (e.g., Friedlander & Robins, 1995; Fraker & Maynard, 1987; Ioannidis et al, 2001). Four randomized experiments we are doing at Johns Hopkins University and the Success for All Foundation illustrate the potential and the pitfalls. One of these, which I mentioned earlier, involves randomly assigning 41 schools to Success for All or control conditions for a 3-year experiment. Initially, we offered

$30,000 to each school, but we got hardly any takers. Schools were unwilling to take a chance on being assigned to the control group for three years.

In spring, 2002, we changed our offer. Schools willing to participate were randomly assigned to use Success for All either in grades K-2 or in 3-5. Recruitment was still difficult, but under this arrangement, we signed up adequate numbers of schools.

For another study led by my colleague Bette Chambers, we recruited schools for a third-party study of the Curiosity Corner preschool model. We offered schools the program for free, to start either in 2003-2004 or 2004-2005 (with random assignment to the two start dates). The 2004-2005 group serves as the control group in 2003-04. This delayed treatment control group design was easy for schools to accept, and we did not have serious recruiting problems. We're doing a nearly identical study of an after-school program, and again, recruitment was not difficult.

We recently completed a study of the use of embedded multimedia, video vignettes embedded in beginning reading instruction (Chambers et al., 2004). Again, ten schools were randomly assigned to receive the multimedia materials immediately or one year later. Finally, my colleague Geoff Borman did randomized evaluations of summer school programs, in which individual children were randomly assigned to participate now or later (Borman, Boulay, Kaplan, Rachuba, & Hewes, 2001). In all of these cases, obtaining sufficient volunteers was not difficult.

These examples of a diverse set of research problems illustrate that one way or another, it is usually possible to use random assignment to evaluate educational programs. There is no one formula for randomization, but with enough resources and cooperation from policy makers, random assignment is possible.

Beyond the benefits for reducing selection bias, there is an important *political* reason to prefer randomized over matched studies at this point in history. Because of political developments in the U.S., we have a once in a lifetime opportunity to reverse the "awful reputation" that educational research has among policy makers (Kaestle, 1993; Lagemann, 2002). This is a time when it makes sense to concentrate resources and energies on a set of randomized experiments of impeccable quality and clear policy importance, to demonstrate that such studies can be done. Over the longer run, I believe that a mix of randomized and rigorous matched experiments evaluating educational interventions may be healthier than a steady diet of randomized experiments, but right now we need to establish the highest possible standard of evidence, on a par with standards in other fields, to demonstrate what educational research can accomplish.

<LEVEL 4> Non-Experimental Research

I should hasten to say again that forms of research other than experiments, whether randomized or matched, can also be of great value. Correlational and descriptive research are essential in theory building and in suggesting variables worthy of inclusion in experiments. Our Success for All program, for example, owes a great deal to correlational and descriptive process-product studies of the 1970's and 1980's (see Slavin & Madden, 2001). As components of experiments, correlational and descriptive studies can also be essential in exploring variables that go beyond overall program impacts. In some policy contexts, experiments are impossible, and well-designed correlational or descriptive studies may be sufficient.

The experiment, however, is the design of choice for studies that seek to make causal conclusions, and particularly for evaluations of educational innovations.

<LEVEL 1> Basing Educational Policy on Evidence

Historically, the impact of education research on education practice has been tenuous at best. Innovation takes place, but it is based on fads and politics rather than evidence. At best, education policies are said to be "based on" scientific evidence, but are rarely scientifically evaluated. This distinction is critical. The fact that a program is based on scientific research does not mean that it is in fact effective. For example, imagine an instructional program whose materials are thoroughly based on scientific research, but which is so difficult to implement that in practice, teachers do a poor job of it, or which is so boring that students don't pay attention, or which provides so little or such poor professional development that teachers do not change their instructional practices. Before the Wright brothers, many inventors launched airplanes that were based on exactly the same "scientifically-based aviation research" as the Wright brothers used at Kitty Hawk, but the other airplanes never got off the ground. Worse, any program or policy can find some research somewhere that suggests it might work.

Given the current state of research on replicable programs in education, it would be difficult to require that government funds be limited to programs that have been rigorously evaluated, because there are so few such programs. However, programs that do have strong, rigorous evidence of effectiveness should be emphasized over those that are only based on valid principles, and there needs to be a strong effort to invest in development and evaluation of replicable programs in every area, so that eventually legislation can focus not on programs "*based* on scientifically-based research" but on programs that have actually been successfully evaluated in rigorous experiments.

<LEVEL 3> Research Syntheses

The evidence-based policy movement is by no means certain to succeed. Education has a long tradition of ignoring or even attacking rigorous research. Researchers themselves, even those who fundamentally agree on methodologies and basic principles, may disagree publicly about the findings of research. These disagreements, which are a healthy and necessary part of the scientific process, will be seized upon by individuals who oppose the entire concept of evidence-based reform as indications that even the experts disagree.

For these and many other reasons, it is essential that independent review commissions representing diverse viewpoints be frequently constituted to review the research and produce consensus on what works, in language that all educators can access. In the area of reading, it is impossible to overstate the policy impact of the National Research Council (1995) and National Reading Panel (Snow, Burns, & Griffin, 1998) reports, which produced remarkable consensus on the state of the evidence in early literacy. Consensus panels of this kind, with deep and talented staff support, should be in operation continually, on a broad range of policy-relevant questions, so that practitioners and policy makers can have a way to cut through all the competing claims and isolated research findings to get to the big picture findings that methodologically sophisticated researchers can agree to represent the evidence fairly and completely. The federally-funded What Works Clearinghouse is carrying out rigorous reviews of research on a range of programs and practices. This effort is just getting under way, but it could become very influential if it gives government funders a basis for favoring well-evaluated practices.

<LEVEL 3> Potential Impact of Evidence-Based Policies on Educational Research

Up to now, I've spoken primarily about the potential impact of evidence-based policies on education policies and practice. I'd now like to consider the potential impact on educational research.

I believe that if evidence-based policies take hold, this will be enormously beneficial for all of educational research, not just research involving randomized or matched experiments. First, I am confident that when policymakers perceive that educational R&D is actually producing programs that are shown in rigorous experiments to improve student outcomes, they will fund research at far higher levels. This should not be a zero-sum game, in which new funds for experiments will be taken from the very limited funds now available for educational research (see Shavelson & Towne, 2002). Rather, I believe that making research relevant and important to policymakers will make them more, not less, willing to invest in all forms of disciplined inquiry in education, be it correlational, descriptive, ethnographic, or otherwise. The popularity of medical research depends totally on its ability to cure or prevent diseases, but because randomized experiments routinely identify effective treatments (and protect us from ineffective treatments), there is vast funding for *basic* research in medicine, including epidemiological, correlational, and descriptive studies. Researchers and developers will be able to argue convincingly that basic research is essential to tell us what kinds of educational programs are worth evaluating.

A climate favorable to evidence-based reform will be one in which individual researchers working on basic problems of teaching and learning will be encouraged and funded to take their findings from the laboratory or the small-scale experiment, or from the observation or interview protocol, to themselves develop and then rigorously evaluate educational treatments. Education is an applied field. Research in education should ultimately have something to do with improving outcomes for children.

<LEVEL 1> Conclusion

Evidence-based policies have great potential to transform the practice of education, as well as research in education. Evidence-based policies could finally set education on the path toward progressive improvement that most successful parts of our economy and society embarked upon a century ago. With a robust R&D enterprise and government policies demanding solid evidence of effectiveness behind programs and practices in our schools, we could see genuine, generational progress instead of the usual pendulum swings of opinion and fashion.

This is an exciting time for educational research and reform. We have an unprecedented opportunity to make research matter, and to then establish once and for all the importance of consistent and liberal support for high-quality research. Whatever their methodological or political orientations, educational researchers should support the movement toward evidence-based policies, and then set to work to generate the evidence that will be needed to create the schools our children deserve.

References

Angrist, J. D. (2004). American education research changes tack. *Oxford Review of Economic Policy, 20* (2), 198-212.

Borman, G., Boulay, M., Kaplan, J., Rachuba, L., & Hewes, G. (2001). *Randomized evaluation of a multi-year summer program: Teach Baltimore.* Baltimore, MD: Johns Hopkins University, Center for Research on the Education of Students Placed at Risk.

Boruch, R.F. (1997). *Randomized experiments for planning and evaluation: A practical guide.* Thousand Oaks, CA: Corwin.

Chambers, B., Cheung, A., Gifford, R., Madden, N., & Slavin, R. E. (2004). *Achievement effects of embedded multimedia in a Success for All reading program.* Baltimore, MD: Success for All Foundation.

Doll, R. (1998). Controlled trials: The 1948 watershed. *British Medical Journal, 317*, 1217-1220.

Fraker, T., & Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources, 22* (2), 194-227.

Friedlander, D., & Robins, P.K. (1995). Evaluating program evaluations: New evidence on commonly used nonexperimental methods. *American Economic Review, 85* (4), 923-937.

Gallagher, J. (2002). What next for OERI? *Education Week, 21* (28), 52.

Herman, R. (1999). *An educator's guide to schoolwide reform.* Arlington, VA: Educational Research Service.

Ioannidis, J.P.A. et al. (2001). Comparison of evidence of treatment effects in randomized and nonrandomized studies. *Journal of the American Medical Association, 286* (7), 821-830.

Kaestle, C.F. (1993). The awful reputation of educational research. *Educational Researcher, 22* (1), 23-26-31.

Lagemann, E.C. (2002, April). *An elusive science: The troubling history of education research.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Levin, M.E., & Levin, J.R. (1990). Scientific mnemonics: Methods for maximizing more than memory. *American Educational Research Journal, 27*, 301-321.

Linn, R.L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis, 24* (1), 29-36.

Mosteller, F., & Boruch, R. (Eds.) (2002). *Evidence matters: Randomized trials in education research*. Washington, DC: Brookings.

Rhine, W.R. (1981). *Making schools more effective: New directions from Follow Through.* New York: Academic Press.

Shavelson, R.J., & Towne, L. (Eds.) (2002). *Scientific research in education.* Washington, DC: National Academy Press.

Slavin, R.E. (1989). PET and the pendulum: Faddism in education and how to stop it. *Phi Delta Kappan, 70*, 752-758.

Slavin, R.E. (1997). Design competitions: A proposal for a new Federal role in educational research and development. *Educational Researcher, 26* (1), 22-28.

Slavin, R.E. (2003). *Educational psychology: Theory into practice.* (7th ed.). Boston: Allyn & Bacon.

Slavin, R.E., & Madden, N.A. (Eds.) (2001). *One million children: Success for All.* Thousand Oaks, CA: Corwin.

Snow, C.E., Burns, S.M., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

Southwest Educational Research Laboratory (2002). CSRD database of schools (web site: http://www.sedl.org/csrd/awards.html) Austin, TX: Author

U.S. Congress (2001). *No Child Left Behind Act of 2001.* Washington, DC: Author.

U.S. Department of Education (2002). *2002 application for new grants: Preschool curriculum evaluation research grant program* (CFDA No. 84.305J). Washington, DC: Author.

U.S. Department of Education (2002b). *Strategic plan, 2002-2007.* Washington, DC: Author.

U.S. Department of Education (2002c). *Draft guidance on the Comprehensive School Reform Program* (June 14, 2002 update). Washington, DC: Author.

Whitehurst, G. (2002). *Charting a new course for the U.S. Office of Educational Research and Improvement.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

The Success for All Middle School: Adding Content to Middle Grades Reform

Cecelia Daniels
Nancy A. Madden
Success for All Foundation

Robert E. Slavin
Johns Hopkins University

Revised, January, 2005

This research was funded primarily by the Institute of Education Sciences, U.S. Department of Education (No. ED-99-CO-0153). However, any opinions expressed are those of the authors and do not necessarily represent IES positions or policies.

Sixteen years ago, the Carnegie Corporation (1989) issued a report that profoundly affected the education of young adolescents. *Turning Points* critiqued the rigid traditional structure of middle schools and advocated reforms intended to make middle grades education more personalized, supportive, and active: interdisciplinary teams, cooperative learning, involvement with families and community, mentoring, and active teaching.

Today, the middle school movement is itself at a turning point. The *Turning Points* reforms, where they have been implemented, have created more humane, child-focused environments that are more in tune with the developmental needs of young adolescents. Yet the achievement of children in the middle grades, especially in high-poverty communities, has continued to languish.

As recognized by one of the intellectual leaders of *Turning Points*, Tony Jackson (Jackson & Davis, 2000), what *Turning Points* and other middle grades reforms of the 1980's and '90's left out was curriculum and instruction. Except for general suggestions about the benefits of active, hands-on, cooperative learning and teaching, teachers have had few practical tools to translate the good ideas of the middle school movement into day-to-day teaching. In the current environment increasing accountability pressures brought on in part by No Child Left Behind, it is simply not enough to engage and support young adolescents. They also need to learn more. Reform in the middle grades needs to incorporate the advances advocated by *Turning Points*, but also to develop well-designed, replicable models that provide challenging content, research-based instructional strategies, and extensive professional development for teachers to enable these students to make progress on the standards that all states are mandating.

The Success for All Middle School

The Success for All Middle School was designed to help middle grades educators implement the most important elements of *Turning Points* and to add well-structured curricula, instructional methods, and professional development for teachers to help students reach their full potential. The program is based on the Success for All elementary design, the most widely used and extensively evaluated of all comprehensive school reform models (Borman et al., 2003; Herman, 1999; Slavin & Madden, 2001). However, the elementary model was totally redesigned to meet the very different developmental needs of young adolescents and the institutional realities of middle grades education. It provides teachers with specific, well-structured student materials, manuals, and other supports, as well as extensive professional development, follow-up, and opportunities for continuing growth.

School and Classroom Organization

Interdisciplinary teams. Like many other middle school reform models, students in Success for All middle schools are grouped in interdisciplinary "teams," each of which has one teacher of each subject. The purpose of these teams is to provide students with a smaller core group of peers and caring adults to attend to their academic and social needs.

Facilitator. Each Success for All Middle School has a full-time facilitator who helps all teachers implement the program, visits classes, organizes data for grouping, and maintains coordination among all staff.

Grouping for Reading. Ensuring literacy for all students is a primary goal of the Success for All Middle School. Students in each grade, 6-8, are assigned to a reading class according to their level of reading skill. A common time period is set aside for this purpose, and all teachers, including art, music, physical education, and other special subject teachers, teach a reading class. Because of this common reading period, students who make rapid progress can be easily moved at any time to higher-performing reading classes without upsetting their entire class schedules. Further, teaching reading gives all teachers strategies in their subject-matter teaching that continuously reinforce literacy skills.

Cooperative Learning. Cooperative learning is extensively used in all subjects in Success for All Middle Schools. Research on cooperative learning has long established that students who work in small, well-structured learning teams gain academically if there are clear group goals and if group success depends on the individual learning of all group members (see Slavin, 1995; Slavin, Hurley, & Chamberlain, 2003). A cooperative group typically involves four students who are diverse in skills, gender, and ethnicity. Students work together on projects and academic work and help each other learn content, but ultimately each student must show individual mastery of the content. Use of cooperative teams also contributes to outcomes such as improved social acceptance, intergroup relations, and self-esteem, all of which are of particular importance for young adolescents (Slavin, 1995).

Curriculum Components

The Reading Edge. The most important curriculum focus of the Success for All Middle School is reading. Reading performance in high-poverty middle schools is unacceptably low (see Donahue et al., 1999; Cooney, 1998), and this deficit holds back progress in all subject areas (Jackson & Davis, 2000).

As students beginning middle school face more challenging content in various subject areas, advanced reading skills and strategies become essential. The Reading Edge program meets this need with an accelerated 60-minute block every day, providing students at all reading levels with structured lessons. Phonics, fluency, vocabulary, and basic and advanced comprehension strategies form the program's foundation. Students learn to understand expository as well as narrative texts and to build the study strategies for success in high school. In addition, Reading Edge lessons make extensive use of cooperative learning, harnessing the strength of peer relationships in young adolescents and giving students powerful incentives to read and to help their teammates read. These reading and cooperative learning techniques are reinforced throughout the day in the other components of the SFA Middle School.

Assessment, Grouping, and Regrouping. At the beginning of the school year, a standardized assessment provides baseline data on each student's reading level (from pre-primer to eighth grade). The SFA facilitator compiles this data to help him or her place students in instructionally appropriate reading levels. As noted earlier, all faculty members teach reading to maximize the number of classes and reduce class size. Having homogeneous classes limits the range of performance levels and allows teachers to customize instruction for individual learning styles. Every eight weeks, students are reassessed and regrouped according to the progress they have made. In this way, students have the opportunity to move more than one reading level each year until they are proficient, on-grade-level readers.

Humanities. The SFA Middle School humanities curriculum challenges students to make practical use of reading, writing, and analytical skills. These standards-based units are taught daily, usually in a double period, and include both social studies and language arts curricula. In social studies units, students investigate important themes and topics, connecting what they learn about the past with their own lives, and present their findings in various forms of writing. In language arts units students explore different genres of literature, write original pieces following conventions of writing, and learn and practice basic grammar skills.

Each grade level begins with one or more foundation units that familiarize teachers and students with cooperative learning techniques and focus on concrete skills that students apply throughout the year. For example, in a foundation unit on the conventions of writing, students learn the steps of the writing process and peer review as they practice working as members of a cooperative learning team. Students then refine their use of the writing process in every unit that follows.

The remaining units engage students with a problem to solve or a task to complete related to a particular theme or topic. For example, a unit on Ancient Egypt challenges students to solve the mystery of a tomb robbery. To do so, they take on the roles of possible suspects from the ranks of Egyptian society. As students learn about life in ancient Egypt, they make decisions about the robbery based on their findings. Such materials engage students' curiosity, emotions, and intellects, enhancing their motivations to learn the content.

Science. Many Success for All Middle Schools use the specially developed science program a year or two after they begin the reading program. In it, students construct knowledge on the basis of direct experience through exploration, teacher demonstration and explanation, and direct instruction and experimentation. All units are based on National Science Education Standards.

Each grade level begins with one or more foundation units that focus on a specific set of skills. For instance, a unit on science safety teaches students not only how to work safely in a science lab, but also how to respond to the classroom management strategies used in SFA classes.

The remaining units present students with a scenario or problem. For example, in Earthquake!, about a fictitious town situated on a fault, students compile recommendations concerning land use, earthquake-resistant building designs, and other issues impacted by seismic activity. In the context of this work, students learn about using models to study earth science concepts such as plate tectonics, as well as physical science concepts such as wave structure and energy. Students also learn to read maps, informational text, organize data into charts and graphs, draw conclusions, and write their findings in a number of different formats.

School and Family Success

School and Family Success teams within each school focus on issues such as attendance, school-based intervention for struggling students, family involvement, service integration with community agencies, and building students' social problem-solving skills.

Evaluation

The Success for All Middle School is being evaluated by a third party evaluator, the National Opinion Research Center (NORC) at the University of Chicago, which is collecting student-level data from state assessments. However, reading results at the school level from 2001 to 2004 were obtained from state web sites. School-level results compare achievement gains on state high-stakes reading measures in SFA middle schools to those in matched comparison schools.

In all seven school pairs, students in the SFA Middle Schools gained substantially more on their state reading assessments than did students in comparison schools. In many cases, these differences were striking. At Tahola School, a K-12 school primarily serving Native American students in rural

Washington State, the Success for All seventh graders gained 95.5 percentage points in students meeting standards on the Washington Assessment of Student Learning (WASL), going from 4.5% to 100% passing. The comparison school gained only 18.4 percentage points, while the state average gained only 20.7 percentage points. Similarly, seventh graders at Richards Middle School in rural Missouri gained 31.5 percentage points in students passing the Missouri Assessment Program (MAP) Reading Scale, while a matched control school gained 10.3 points and the state gained only 2.4 points. Two inner-city middle schools in Indianapolis gained markedly more than their comparison schools. Coleman Middle school gained 9.0 percentage points on the Indiana Statewide Testing for Educational Progress (ISTEP), averaging grades 6 and 8. A control school gained only 0.5 points. Longfellow Middle School gained 15.5 percentage points on ISTEP, while its control school gained 4.0 points. Indiana middle grades as a whole gained 7.0 percentage points. Carver Middle School in Meridian, Mississippi gained 5.8 percentage points in students passing the Mississippi Curriculum Test (MCT), while its control school gained by 2.3 points. SFA middle schools in Arizona and Louisiana gained on their state assessments while both their control schools and their states lost ground.

Recognizing the problems inherent in averaging across different state measures, it is still interesting to note that across the seven SFA schools, students gained an average of 24.6 percentage points on state reading tests, while matched control schools gained only 2.2 points and the gain of in each of the schools' respective states was 4.2 percentage points.

Table 1 summarizes the gains in each SFA school, its matched control, and its state.

Table 1
Gains in Percent of Students Passing State Reading Tests in Success for All and Control Middle Schools, 2001 to 2004

			Gains in Percent Passing		
School (State)	Measure	Grades Tested	SFA	Control	State
Washington	WASL	7	+95.5	+18.4	+20.7
Missouri	MAP	7	+31.5	+10.3	+2.4
Indiana-pair 1	ISTEP	6, 8	+9.0	+0.5	+7.0
Indiana- pair 2	ISTEP	6, 8	+15.5	+4.0	+7.0
Mississippi	MCT	6, 7	+5.8	+2.3	+8.1
Arizona	AIMS	8	+3.0	-12.0	-6.0
Louisiana	LEAP	8	+12.0	-8.0	-5.0
Means*			+24.6	+2.2	+4.2

*Means across different state assessments should be interpreted cautiously.

Conclusion

The Success for All Middle School design is a comprehensive, replicable model for middle schools serving many at-risk young adolescents. Not only does it incorporate the structural features emphasized in *Turning Points*, but it also goes beyond this to provide specific content, instructional strategies, and professional development to help all teachers implement state-of-the-art instruction in their classes. Third-year evaluation data show that this approach is having a substantial impact on students' reading achievement in all of its pilot schools. As the Success for All Middle School and other content-focused middle school reforms begin to work at a larger scale and continue to produce convincing data, we may finally achieve the breakthrough that *Turning Points* promised sixteen years ago: reliable, replicable models to help schools ensure the success of young adolescents.

References

Borman, G.D., Hewes, G.M., Overman, L.T., & Brown, S. (2003) Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73* (2), 125-230.

Carnegie Council on Adolescent Development (1989). *Turning Points: Preparing American Youth for the 21st Century.* New York: Carnegie Corporation.

Cooney, S. (1998). *Education's weak link: Student performance in the middle grades.* Atlanta: Southern Regional Education Board.

Daniels, C., Chamberlain, A., Madden, N., & Slavin, R. (2003). *The Success for All Middle Schools: Findings of early evaluations.* Baltimore, MD: Success for All Foundation.

Donahue, P.L., Voelkl, K.E., Campbell, J.R., & Mazzeo, J. (1999). *NAEP 1998 reading report card for the nation.* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Herman, R. (1999). *An educator's guide to schoolwide reform.* Arlington, VA: Educational Research Service.

Jackson, A.W., & Davis, G.A. (2000). *Turning points 2000: Educating adolescents in the 21st century.* New York: Teacher's College Press.

Slavin, R.E. (1995). *Cooperative learning: Theory, research, and practice* (2nd Edition). Boston: Allyn & Bacon.

Slavin, R.E., & Madden, N.A. (Eds.) (2001). *One million children: Success for All.* Thousand Oaks, CA: Corwin.

Slavin, R.E., Hurley, E.A., & Chamberlain, A.M. (2003). Cooperative learning and achievement: Theory and research. In W.M. Reynolds & G.E. Miller (Eds.), *Handbook of Psychology*, Volume 7 (pp. 177-198). Hoboken, NJ: Wiley.

Scientifically Based Research: A Reader's Guide

Robert E. Slavin
Johns Hopkins University

August, 2002

This article was written under funding from the Office of Educational Research and Improvement, U.S. Department of Education (Grant No. OERI-R-117-D40005). However, any opinions expressed are those of the author and do not necessarily represent OERI positions or policies.

In every successful, dynamic part of our economy, *evidence* is the force driving change. In medicine, medications or procedures are constantly being developed, evaluated in comparison to current practice, and if they produce greater benefits, disseminated widely. In agriculture, better seeds, better equipment, and better farming methods are developed, tested, and then eagerly adopted. In technology, in engineering, in field after field, progress comes from R&D. Physicians, farmers, consumers, and government base key decisions on the results of rigorous research.

In education, on the other hand, research has played a relatively minor role in reform. Untested innovations appear, are widely embraced, and then disappear as their unrealistic claims fail to materialize. They are then replaced by equally untested innovations diametrically opposed in philosophy, in an endless swing of the reform pendulum. Far more testing goes into our children's hair gel and acne cream than into most of the curricula or instructional methods used by their teachers. Yet which of these is more important to our children's future?

Evidence-Based Reform

At long last, education reform may be entering an era of progress based on implementation of well-researched programs and practices (see Slavin, in press). There is a new interest in government in the research base for programs adopted by schools. This began with the Comprehensive School Reform Demonstration (CSRD) legislation in 1997, which gives grants to schools to adopt "proven, comprehensive" reform designs. Ideally, "proven" means that programs have been evaluated in "scientifically-based research," involving comparisons to matched or randomly assigned control groups. The Bush Administration's No Child Left Behind Act refers to "scientifically-based research"

110 times with reference to Reading First programs for reading in grades K-3, Early Reading First for pre-K, Title I school improvement programs, and many more. In each case, schools, districts, and states must justify the programs they expect to implement under federal funding. "Scientifically-based research" is defined as "rigorous, systematic, and objective procedures to obtain valid knowledge," with an emphasis on evaluations that use experimental or quasi-experimental designs, preferably with random assignment.

Research Matters

The many policy changes basing education funding and practice on scientifically-based, rigorous research have important consequences for educators. Research matters. Educators have long given lip service to research as a guide to practice. But increasingly, they are being asked to justify their choices of programs and practices using the findings of rigorous, experimental research.

Judging the Validity of Educational Research

Educators who are called upon to use programs based on "scientifically based research" are often uncomfortable making such judgments. Why is one study valid while another is not?

There are many kinds of research done for many reasons, but for evaluating the achievement outcomes of educational programs, judging research quality is relatively straightforward. Valid research for this purpose compares several schools that used a given program to several carefully matched control schools on meaningful measures of achievement. It's that simple.

Control Groups

A hallmark of valid, scientifically-based research on educational programs is the use of control groups. In a good study, several schools using a given program are compared to several schools that are very similar in demographics and prior performance, preferably in the same school district. At least five schools in each group is desirable; studies with just one or two schools in each group can be biased by circumstances unique to a given school.

A control group is intended to provide an estimate of what students in the experimental program would have achieved if they'd been left alone. That's why it's essential that the control schools be as similar as possible to the program schools at the outset.

Randomized and Matched Experiments

The most convincing form of a control group comparison is a randomized experiment, in which students, teachers, or schools are assigned by chance to one group or another. For example, the principals and staffs at ten schools might express interest in using a given program. The schools might be paired up and then assigned by coin flip to the experimental or control group. Randomized experiments are very rare in education, but can be very influential. Perhaps the best known example in recent years is the Tennessee class size study, or STAR (Achilles, Finn, & Bain, 1997/98), in which children were assigned at random to small classes (15 students), regular classes (20-25 children), or regular classes with an aide. The famous Perry Preschool Program (Berrueta-Clement et al., 1984) assigned four-year-olds at random to attend an enriched preschool program or stay at home. Two recent studies of James Comer's School Development Project (SDP) randomly assigned schools to use SDP or keep using their current program (Cook, Hunt, & Murphy, 2000; Cook et al, 1999). In each of these studies, random assignment made it very likely that the experimental and control groups were identical at the outset, so any differences at the end are sure to have been due to the program.

Matched studies are far more common than randomized ones. In a matched program evaluation, students in a given program are compared to those in a control group that is similar in prior achievement, poverty level, demographics, and so on. Matched studies can be valid if the experimental and control groups are very similar. Often, statistical methods are used to "control for" pretest differences between experimental and control groups. This can work if the differences are small, but if there are large differences at pretest, statistical controls or use of gain scores (post minus pre) are generally not adequate.

Problems with Matched Studies

The potential problem with even the best matched studies is the possibility that there are (unmeasured) characteristics of the schools that chose a given program that are different from those that did not choose it. For example, imagine that a researcher asked ten schools to implement a new program. Five enthusiastically take it on, while five refuse. Using the refusal group as a control group, even if it is similar in other ways, can introduce what is called selection bias. In this example selection bias would work in favor of finding a positive treatment effect, because the volunteer schools are likely to have more enthusiastic, energetic teachers willing to try new methods than the control schools. However, in other cases it may be the most desperate or dysfunctional schools that chose or were assigned to a given program, giving an advantage to the control schools.

Is Random Assignment Essential?

Random assignment to experimental and control groups is the gold standard of research. It virtually eliminates selection bias, as students, classes, or schools were assigned to treatments not by their own choice, but by the flip of a coin or another random process.

Because of the ability of randomized studies to rule out selection bias, the U.S. Department of Education and many researchers and policy makers have recently been arguing for a substantial increase in the use of randomized designs in evaluations of educational programs. Already, there are more randomized studies under way in education than at any point in history.

The only problem with random assignment is that it is very difficult and expensive to do, especially for school-wide programs (where random assignment of whole schools is necessary). No one likes to be assigned at random, so such studies often have to provide substantial incentives to get educators to participate. Still, such studies are possible; we have such a study under way to evaluate our Success for All comprehensive reform model, and as noted earlier there were two randomized studies of Comer's School Development Program.

At present, with the movement toward greater use of randomized experiments in education in its infancy, educators evaluating the research base for various programs have to look carefully at well-matched experiments, valuing those that try to minimize bias by using closely matched experimental and control groups, adequate numbers of schools, avoiding comparing volunteers to non-volunteers, and so on. However, this situation may change in coming years.

Statistical and Educational Significance and Sample Size

Reports of educational experiments always report statistics indicating whether or not there is a statistically significant difference between the achievement of children in the experimental group and those in the control group, usually controlling for pretests and/or other factors. A usual criterion is "$p< .05$," which means that the probability is less than 5% that an observed difference might have happened by chance.

If students in a program had "significantly higher" scores than those in a control group, that's important, but it may not be enough. In a very large study, a very small difference could be significant. A typical measure of the size of a program effect is "effect size," the experimental-control difference divided by the control group's standard deviation (a measure of dispersion of scores). In educational experiments, an effect size of +0.20 (20% of a standard deviation) is often considered a minimum for educational significance; effect sizes above +0.50 would be considered very strong.

There is a problem inherent to educational experiments that is very important to be aware of. Children are grouped in classes and schools, and these groupings can have a profound impact on student outcomes. Often, an experiment will compare one school using Program X to one matched control school. If there are 500 children in each school, this is a very large experiment. Yet the difference between the Program X school and the control school could be due to any number of factors that have nothing to do with Program X. Perhaps the Program X school has a better principal or a cohesive group of teachers or has been redistricted to include a higher-performing group of children. Perhaps one of the schools experienced a disaster of some sort—in one early study of our Success for All program, Hurricane Hugo blew the roof off of the one Success for all school but did not affect the one control school.

Because of the possibility that something unusual that applies to an entire school could affect scores for all the children in that school, statisticians demand that statistical analyses be done using *school* means, not individual student scores. This way, any school factors are likely to balance out. This requires at least 20-25 *schools* per condition. Very few educational experiments are this large, however, so the vast majority of experiments analyze at the student level.

Readers of research should apply a reasonable approach to this problem. Studies with a single school or class per condition should be viewed with great caution. However, a study with as many as five program schools and five control schools probably has enough schools to make it unlikely that a single unusual school could skew the results. Such a study would still use individual scores, not school means, but it would be far preferable to a one school to one school comparison.

A single study involving a small number of schools or classes may not be conclusive in itself, but many such studies, preferably done by many researchers in many locations, can add confidence that a program's effects are valid. In fact, this is how experimental research in education usually develops. Rather than one large, definitive study, there are usually many that are small and flawed in various (unbiased) ways, but if they tend to find consistent effects, the entire set of studies may produce a meaningful conclusion even if no one study is conclusive by itself.

Research to Avoid

All too often, program developers or advocates cite evidence that is of little value, or is downright misleading. A rogue's gallery of such research is as follows.

Cherry Picking

Frequently, program developers or marketers report on a single school or a small set of schools that made remarkable gains in a given year. Open any educational magazine and you'll see an ad like this: "Twelfth Street Elementary went from the 20th percentile to the 60th in only one year!" Such claims have no more validity than advertisements for weight loss programs that tell the story of one person who lost 200 pounds (forgetting to mention the perhaps hundreds who did not lose weight on the diet). This kind of "cherry picking" is easy to do in a program that serves many schools, as there are always individual schools that make large gains in a given year, and the marketer can pick them after the fact, just by looking down a column of numbers to find a big gainer. (Critics of the program can use the same technique to find a big *loser*). Such reports are pure puffery, not to be confused with science.

Bottom Fishing

A variant of cherry-picking is "bottom fishing," using an after-the-fact comparison in which an evaluator compares schools using a given program to matched "similar schools" known to have made poor gains in a given year. Comparisons can be appropriately made between gains in program schools and gains made in the entire district or state, as the large, exhaustive comparison group makes "bottom fishing" impossible. However, after-the-fact studies purporting to compare groups selected by the evaluator should be interpreted with caution.

Pre-Post Studies

Another common but misleading design is the pre-post comparison, lacking a control group. Typically, this is used with standardized test data, with the rationale that the expected year-to-year gain in percentiles, normal curve equivalents (similar to percentiles), or percent passing is zero, so any school that gained more than zero is making good progress.

The problem with this logic is that many states and districts make substantial gains in a given year, so the program schools may be doing no better than other schools. In particular, states usually make rapid gains in the years after they adopt a new test. At a minimum, program schools in a given district or state should be compared to the gains made in the entire district or state, as noted earlier.

"Based on Scientifically-Based Research" versus Rigorously Evaluated

A key issue in the recent No Child Left Behind legislation is the distinction between programs that are "based on scientifically-based research" and those that have themselves been evaluated in scientific, valid experiments. A program can be "based on scientifically-based research" if it incorporates the findings of rigorous experimental research. For example, reading programs are eligible for funding under the federal Reading First initiative if states determine that they incorporate a focus on five elements of effective reading instruction: phonemic awareness, phonics, fluency, vocabulary, and comprehension. These elements were identified by the National Reading Panel (1999) as having been established in rigorous research, especially randomized experiments. Yet there is a big difference between a program *based* on such elements and a program that has itself been compared to matched or randomly assigned control groups. It is easy to imagine a reading program that would incorporate the five elements but whose training was so minimal that teachers did not implement these elements well, or whose materials were so boring that children were not motivated to study them, for example. The No Child Left Behind guidance (U.S. Department of Education, 2002) recognizes this distinction, and notes a preference for programs that have been rigorously evaluated. , but also recognizes that requiring such evaluations would screen out many new reading programs that have not been out long enough to have been evaluated. This may make sense from a pragmatic or political perspective, but from a research perspective a program that is unevaluated is unevaluated, whether or not it is "based on" scientifically-based research. A basis in scientifically-based research makes a program promising, but not proven.

Research Reviews

In order to judge the research base for a given program, it is not necessary that every teacher, principal, or superintendent carry out his or her own review of the literature. Several reviews applying standards much like those in this article have summarized evidence on various programs.

For comprehensive school reform (CSR) models, there are several useful reviews. The American Institutes for Research (AIR) carried out a review of 24 CSR programs (Herman, 1999), taking a hard-nosed look at the evidence behind the most widely used CSR programs. The Thomas Fordham Foundation (Traub, 1999) commissioned an evaluation of ten popular CSR models. Berman, Hewes, Rachuba, & Brown (2002) carried out a meta-analysis (or quantitative synthesis) of research on 29 CSR models.

The Northwest Regional Laboratory (2002) publishes a regularly updated compendium of comprehensive reform models and single-subject programs that could be incorporated into CSR models. Slavin & Fashola (1998) reviewed research on a wide range of educational innovations, both single-subject and comprehensive models, and Slavin & Calderón (2001) reviewed research on programs used with Latino students. Pearson & Stahl (2002) reviewed research on 21 innovative beginning reading programs.

These reviews facilitate the process of evaluating the evidence behind a broad range of programs, but it's still a good idea to look for a few published examples of studies on a program of interest, to get a sense for the nature and quality of the evidence supporting a given model. Also, it's useful to look at multiple reviews, if possible, as reviewers differ in their conclusions and recommendations (because their review criteria differ). Adopting a program for a single subject, much less for an entire school, requires a great deal of time, money, and work, and can have a profound impact on a school for a long time. It's well worth the time it takes to look at the research evidence with some care before making such an important decision. Accepting the developer's word for a program's research base is not a responsible strategy.

How Evidence-Based Reform Will Transform Our Schools

The movement to ask schools to adopt programs that have been rigorously researched could have a profound impact on the practice of education, and on the outcomes of education for children. If this movement prevails, educators will increasingly be able to choose from among a variety of models known to be effective if well implemented, rather than reinventing (or misinventing) the wheel in every school. There will never be a guarantee that a given program will work in a given school, just as no physician can guarantee that a given treatment will work in every case. However, a focus on rigorously evaluated programs can at least give school staffs confidence that their efforts to implement a new program are likely to pay off in higher student achievement.

In an environment of evidence-based reform, developers and researchers will be continually working to create new models and improve existing ones. As in medical innovation, today's substantial improvement will soon be replaced by something even more effective. Rigorous evaluations will be common, both to replicate evaluations of various models and to discover the conditions necessary to make programs work. Reform organizations will build capacity to serve hundreds or thousands of schools. Educational leaders will become increasingly sophisticated in judging the adequacy of research, and as a result the quality and usefulness of research will grow. In programs such as Title I, governmental support will increasingly focus on helping schools adopt proven programs, and schools making little progress toward state goals may be required to choose from among a set of proven programs.

Evidence-based reform would finally bring education to the point reached early in the 20th century by medicine, agriculture, and technology, in which evidence is the lifeblood of progress. No Child Left Behind, Reading First, Comprehensive School Reform, and related initiatives have created the possibility that evidence-based reform can be sustained and can become fundamental to the practice of education. Informed educational leaders can contribute to this as well. It is ironic that the field of education, with its value on knowledge, has so embraced ideology rather than knowledge in its own reform process. Evidence-based reform honors the best traditions of our profession, and promises to progressively transform schooling for all children.

References

Achilles, C.M., Finn, J.D., & Bain, H.P. (1997/98). Using class size to reduce the equity gap. *Educational Leadership, 55* (4), 40-43.

Cook, T.D., Habib, F., Phillips, M., Settersten, R.A., Shagle, S., & Degirmencioglu, M. (1999). Comer's school development program in Prince George's County, Maryland: A theory-based evaluation. *American Educational Research Journal 36* (3), 543-597.

Cook, T., Murphy, R.F., & Hunt, H.D. (2000). Comer's school development program in Chicago: A theory-based evaluation. *American Educational Research Journal, 37* (2), 543-597.

Berrueta-Clement, J.R., Schweinhart, L.J., Barnett, W.S., Epstein, A.S., & Weikart, D.P. (1984). *Changed lives.* Ypsilanti, MI: High/Scope.

Borman, G.D., Hewes, G.M., Rachuba, L.T., & Brown, S. (2002). *Comprehensive school reform and student achievement: A meta-analysis.* Submitted for publication. (Request from the author at gborman@education.wisc.edu)

Herman, R. (1999). *An educator's guide to schoolwide reform.* Arlington, VA: Educational Research Service.

National Reading Panel (1999). *Teaching children to read.* Washington, DC: U.S. Department of Education.

Northwest Regional Educational Laboratory (2002). *Catalog of school reform models.* Portland, OR: Author.

Pearson, P.D., & Stahl, S. (2002). *Choosing a reading program: A consumer's guide.* (Technical report). Berkeley, CA: University of California. (Request from the author at ppearson@socrates.berkeley.edu)

Slavin, R.E. (in press). Evidence-based education policies: Transforming educational practice and research. *Educational Leadership.*

Slavin, R.E., & Calderón, M. (Eds.) (2001), *Effective programs for Latino students.* Mahwah, NJ: Erlbaum.

Slavin, R.E. & Fashola, O.S. (1998). *Show me the evidence: Proven and promising programs for America's schools.* Thousand, Oaks, CA: Corwin.

Traub, J. (1999). *Better by design? A consumer's guide to schoolwide reform.* Washington, DC: Thomas Fordham Foundation.

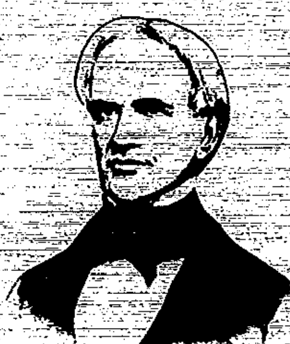
U.S. Department of Education (2002c). *Draft guidance on the Comprehensive School Reform Program* (June 14, 2002 update). Washington, DC: Author.

Presented To

DR. BOB SLAVIN

For Leadership In The Strengthening Of Public Education In The United States

Presented Saturday, February 20, 1999
New Orleans, Louisiana



The purposes of the Horace Mann League are patriotic and educational. Its chief object shall be to foster and strengthen the American public schools and increase the esteem in which the public school is held and the appreciation of its importance among Americans.

The Horace Mann League - Founded In 1922

Robert E. Slavin

1998 JAMES BRYANT CONANT AWARD RECIPIENT

FOR OUTSTANDING CONTRIBUTIONS TO AMERICAN EDUCATION

EDUCATION COMMISSION OF THE STATES

2004

Distinguished Achievement Award

For Excellence in Educational Publishing

FINALIST

DISTINGUISHED ACHIEVEMENT AWARD

Print Publications: Periodicals
Editorial - Learned Article (Adults)

A Reader's Guide to Scientifically Based Research

Educational Leadership, February 1, 2003
The Association for Supervision and Curriculum Development (ASCD)
Robert E. Slavin, Author

Hugh Roome, President

Charlene F. Gaynor, Executive Director

The Palmer O. Johnson Award

In recognition of the lifelong achievement of Palmer O. Johnson as a dedicated educator, and further in recognition of his pioneer work in educational research and methodology his former colleagues and students have founded the Palmer O. Johnson Memorial Award. The Award will be given for an outstanding article selected annually from the American Educational Research Journal, Review of Educational Research, Journal of Educational Statistics, or the Educational Researcher.

Presented to

Robert B. Slavin

for his article Best Evidence Synthesis: An Alternative to Meta-Analytic and Traditional Reviews

Amissville Virginia