



# Planning and Evaluating Change at Scale: Lessons From Reading First

Michael C. McKenna and Sharon Walpole

The evaluation of Reading First, the U.S. Department of Education's multibillion-dollar K–3 initiative, although flawed, nevertheless offers instructive guidance for gauging the impact of future initiatives. After providing an overview of the program, its evaluation, and the historical context of federal initiatives, the authors outline limitations in applying scientific principles at scale. They argue for more nuanced approaches, including meta-analyses across projects, the use of improved statistical approaches, and the incorporation of formative designs. They conclude with four recommendations for evaluating future initiatives. Such evaluations should (a) account for fidelity systematically, (b) include outcome measures that gauge school climate and administrative support, (c) include multiple designs and aggregate the results, and (d) account for the length of implementation.

**Keywords:** federal programs; literacy policy; program evaluation; reading; Reading First

**R**eading First, at a cost of nearly a billion dollars per year over its 8-year duration, is by far the most ambitious federal reform initiative ever undertaken in the area of reading. The decision to discontinue funding, based in part on the results of a contentious national evaluation (Stern, 2008), raises important questions about how the success of such initiatives is gauged. Our purpose in this article is to examine these issues and to distill lessons that may better inform the manner in which we judge the impact of future large-scale programs.

Reading First, a part of the No Child Left Behind Act, provided funds to states and territories to guide changes in K–3 instruction in their most troubled elementary schools. The use of the funds was limited principally to assessments, instructional materials, and professional development. These choices had to reflect a specific body of research on early reading development, embodied in the findings of the National Reading Panel (National Institute of Child Health and Human Development [NICHD], 2000), findings that proved somewhat controversial because of the strict methodological guidelines that the panel adopted in the selection of studies. Once enacted, the funding provisions occasioned intense marketing on the part of commercial enterprises,

each claiming that its wares were aligned with the provisions of the legislation and, consequently, with the findings of the panel (Stern, 2008). Most of the schools with Reading First funding did, in fact, purchase expensive new reading materials and assessments. They also engaged in layers of professional development, from large-scale teacher trainings to on-site, ongoing coaching. Theoretically, this combination of new resources, new assessments to guide instructional decisions, and new training of and support for teachers should have resulted in improved achievement for students. Theoretically, the efficacy question should have been fairly simple to answer.

## Distinguishing Misconceptions From Reality

Three factors complicated the process of evaluating Reading First, and it is important to acknowledge them from the outset. The first, and perhaps the most damaging, involved inappropriate exercise of influence at the outset of the program. It is not surprising that legislation of this scope would be influenced by the views of specific individuals and groups (McDaniel, Sims, & Miskel, 2001; Miskel & Song, 2004). Even before the start of Reading First, critics charged that the federal government had deliberately distorted the panel's findings in an effort to shape a national reading curriculum that was envisioned by a few insiders (Allington, 2002). Their suspicions were, in part at least, validated by the Inspector General's report, which documented malfeasance on the part of a limited number of individuals connected with Reading First (Office of Inspector General, 2006). In fact, the claim that some individuals developing the legislation and guidance for the program favored specific reading programs—rather than the concept of scientifically based instruction—is difficult to refute (e.g., Roller, 2009). For example, 46 states elected to use a single assessment system, the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), after it had been inappropriately singled out in technical assistance sessions sponsored by the U.S. Department of Education (Higgins, 2007). Such facts are unfortunate because they make a reasoned appraisal of the program itself far more difficult (Stern, 2008). In attempting such an appraisal, however, it is important to keep these scandalous actions in proper perspective, for although they influenced state applications for funding, they had little impact on districts or schools; and what influence they did have waned over time as many successful programs were enacted (International Reading Association, 2006).

The second complication for evaluation lies in challenging widespread misconceptions about the nature of the program as it was actually implemented. Our interaction with colleagues who have not had a direct association with Reading First has revealed a number of misconceptions about what was actually required of schools and teachers. These notions, perhaps derived from incidental and anecdotal sources, include the belief that Reading First required that all schools adopt scripted core programs and implement the same well-delineated plan. Ironically, if these assumptions had been accurate, evaluating Reading First would have been more straightforward. The reality is that states interpreted the requirements differently, and in most cases wide latitude was afforded to schools that adopted the program, so long as they maintained a protected literacy block, used core and supplemental materials that embraced scientifically based instructional methods, provided comprehensive professional development, and conducted systematic assessments to inform instruction. And, as Pressley (2005) correctly observed, Reading First guidance was silent about certain factors that are clearly instrumental to effective instruction, such as motivation and classroom management; schools were free to make their own decisions regarding improvements in these areas. Thus, from a design and evaluation standpoint, the limited constraints imposed by Reading First resulted in broad variability in implementation across states, districts, and schools. Reading First schools were not so amorphous as to be unrecognizable, however. The constraints, though few in number, constituted a considerable reform in the target schools.

Finally, there was extensive variance in the assessments that states used to provide summative evaluation. All states designed assessment plans that included screening, diagnosis, progress monitoring, and outcome measures (Kame'enui et al., 2006). DIBELS assessments dominated the program in terms of the first three elements, but state choices varied with respect to outcome measures. Some used their state criterion-referenced test, some used independent standardized tests, and some used subtests of the DIBELS battery. This variance is consistent with the tradition of state control of education standards, reflected in differences in NAEP performance and fueling the current move to a common set of standards and assessments (Duncan, 2009). In addition, states' interpretations of scores varied; the requirement that they measure increases in the percentages of children reading at grade level meant that they constructed definitions of the phrase *at grade level* with reference to specific scores on tests not designed for that purpose. For example, states using norm-referenced tests may have used scores corresponding to the 30th, 40th, or 50th percentile rank to define that standard—very different interpretations indeed. Thus the question of efficacy was nested within states from the start, and the states had very different plans and standards.

### **The Tender Trap of Science**

Although evaluation of the efficacy of Reading First, in particular, was complicated by scandal, by the diversity of the program as it was actually enacted, and by the available achievement data, the experience speaks to a larger issue for education researchers—the hegemony of scientific methodologies. The ascendance of scientifically based research occurred largely in reaction to the perceived

inadequacies of alternative forms of inquiry for addressing issues of efficacy. An example specific to literacy inquiry is the Society for the Scientific Study of Reading, founded in 1992 by Ronald P. Carver as a reaction against the approaches to research favored by proponents of whole language (Carver, personal communication, 1991; see also Edelsky, 1990). In general, however, the narrowing of what constituted scientific evidence was gradual and driven by policy. Eisenhart and Towne's (2003) chronology of the changing definitions of the term *scientifically based* reveals much about the complex, uneasy marriage of research and policy that drove the evolution of the term. In legislative documents alone, what constitutes such research has morphed from a broad commitment to rigorous empirical work (including both quantitative and qualitative research studies) to one that favors only experimental or quasi-experimental designs (Eisenhart & Towne, 2003).

This trajectory has occurred at many levels. Over the past half century, the trend in federal program evaluation has increasingly favored experimental designs (King, 2003). In fact, the National Center for Education Evaluation and Regional Assistance (NCEE) makes clear that scientific paradigms are the mainstay of such evaluations, stating on its website that it “designs evaluation studies to produce rigorous scientific evidence on the effectiveness of education programs and practices” (NCEE, n.d., “NCEE Evaluation Studies”). This trend in evaluation is merely the endpoint of the five-goal, stair-step model espoused by the Institute of Education Sciences. In that model, scientific paradigms are the method of choice, from the development of a promising approach to full scale-up. Indeed, it is probably natural to expect a program grounded in scientific research to be judged by the same standards. However, there are many difficulties with this reasoning, and a frank consideration may help to prevent recurrences of ill-advised evaluations and their inevitable consequence, ill-informed policy.

The principal hazard is that applying scientific principles at scale sacrifices important understandings that affect generalizability. Much is lost by strict adherence to the what-works dichotomy, and evaluations that attend to the conditions under which a program is effective are of far more use to policy makers than the reductionist bottom line reached by approaches that, while statistically elegant, ignore nuances that may be explanatory. The study of the effects of a new approach to instruction, where children are the unit of analysis, will result in an ultimate test of significance, to be sure. However, the approach may prove more effective with some students than others, and a good design will enable the investigator to identify the characteristics of students. At the program level, where schools are likely to be the unit of analysis, the evaluation may well be limited to a global determination of impact; or, worse, nuanced findings regarding the kinds of schools affected may be ignored by policy makers.

Generalizing across schools is a dangerous game. Schools are exceedingly complex systems with intricate combinations of individual characteristics that are easily traced to achievement differences. Schools differ in their student composition (e.g., class size, socioeconomic status, ethnicity, home language), and teachers differ in their knowledge and dispositions. In addition, the schools most in need of new ideas are most likely to be structurally impoverished—to have a larger proportion of unqualified teachers, higher teacher turnover, and lower student achievement (Darling-Hammond, 2006). The same structural difficulties make large-scale

experiments very difficult to accomplish in such settings. Ignoring these differences is simply not good science. There is a fundamental difference between validating an instructional approach and evaluating a program that incorporates that approach. Validation studies are conducted to inform classroom practice; evaluations are carried out to inform policy. Employing the same tools to accomplish both goals involves inherent limitations.

### **What Was: The Reading First Evaluation Plan**

The issue of generalizability was particularly instructive in the case of Reading First. Because the statute precluded random assignment of schools to the program, matched schools were used in the national evaluation, which employed a regression discontinuity design (Gamse, Jacob, Horst, Boulay, & Unlu, 2008). This was a next-best approach that, despite a credible history, suffered in this case from substantive flaws, such as the diffusion of treatment into comparison schools (Reading First Federal Advisory Committee, 2008). But, for the sake of argument, let us put such issues aside and assume that an incontrovertible national evaluation design had identified little advantage for Reading First schools in terms of standardized achievement measures. Such a finding would have been likely to influence policy makers in deciding the program's fate. And in fact, on the basis of a weaker design and a hint of scandal, this is exactly what happened (Stern, 2008). We argue that even in the hypothetical case of a "perfect" evaluation, an all-or-nothing judgment would be ill-advised.

When we mentioned the problem of treatment diffusion to a colleague—a skeptic of Reading First on philosophical grounds—his response was unsympathetic. "Live by the test, die by the test," he told us. Our initial view was that this judgment would be justifiable had the design flaws been remedied. Certainly, the position of the Institute of Education Sciences has been that applying scientific approaches at scale is, in principle, no different from applying them in smaller, tightly controlled settings, even when only a handful of participants are involved (see Slavin, 2008). The shortsightedness of this reasoning lies in its implications for policy. After investing billions, were not stakeholders entitled to a more fine-grained analysis, not simply of *whether* but also of *where* and *how well* the program worked? We argue that to continue adhering to a simplistic thumbs-up-thumbs-down mentality is to risk losing the opportunity to improve programs that show promise. This is not to say that poor programs should be continued indefinitely as investigators search for redeeming merits. But no fewer than 6% of American elementary schools were involved in the Reading First endeavor; from a project this massive, it is reasonable to suggest that far more might have been learned.

Reading First is the perfect poster child for the need to employ formative approaches to the evaluation of large-scale initiatives. The fact that 82% of the states reported that the program was positively affecting student achievement (Center on Education Policy, 2006, 2007) is hard to reconcile with the official evaluation. There are three principal reasons for the disparate findings at the state and school levels. The first is that wide latitude was granted to states in crafting their plans, so long as they remained within the modest parameters we have mentioned. Although it is fair to argue that this freedom was an explicit part of the program and should not be used to excuse its ineffectiveness, it was nevertheless the impetus for a variety worth exploring. The second reason, related

to the first, is that the designers of Reading First combined the findings of the National Reading Panel (NICHD, 2000) in ways that were untested in practice (Pressley, 2005; Roller, 2009). The panel synthesized evidence concerning specific instructional practices but was silent about the most effective ways of combining them into a coherent program. Imagine a recipe that called for mixing, in untested proportions, a number of ingredients known to taste good individually. Finally, although each state was required to identify outcome measures, there was no uniformity across the states in the assessments selected, nor was any state required to participate in the national evaluation. This mistake was not made in the evaluation of the sister program, Early Reading First (Jackson et al., 2007). Further complicating the situation was the fact that state-level evaluations varied considerably in design and even in the questions they sought to answer.

### **What Might Have Been: The Potential for Rigorous and Useful Evaluation**

Given these circumstances, how might the evaluation have been carried out more productively? More important, how might future evaluations be planned from the outset to account for, and in fact capitalize on, inevitable variations in how a program is instantiated in schools and classrooms? Our suggestions combine calls from multiple sources.

#### *1. Account for fidelity systematically.*

At a minimum, levels of implementation must be gauged to determine whether impact varies with respect to fidelity and, if so, what factors have proved conducive to higher levels of implementation (Hamilton, McCaffrey, Stecher, Klein, Robyn, & Bugliari, 2003; Stein et al., 2008). The results of a fidelity-sensitive evaluation could fuel a recursive process in which incremental improvements are systematically made. Slavin (2002) has suggested "an ascending spiral: rigorous research demonstrating positive effects of replicable programs on important student outcomes would lead to increased funding for such research, which would lead to more and better research and therefore more funding" (p. 17). The advent of formative experiments in the realm of established methodologies (Reinking & Bradley, 2004) lays the foundation for applying these ideas on a larger scale; indeed, the Reading for Understanding Research Initiative of the Institute of Education Sciences calls for proposals to design and test interventions using an iterative design so that the interventions can be refined as soon as implementation and effectiveness data are available (Institute of Education Sciences, n.d., "Reading for Understanding Research Initiative"). A consequence of this approach is that the nature of fidelity will change with each formative alteration to the program, but such change reflects the potential for organic growth. This perspective is at odds with the conventional notion that the more specific a program, the greater its chances of having the intended effect (Porter, Floden, Freeman, Schmidt, & Schwille, 1988). Desimone (2002), in supporting the latter notion, observes that "the factors that increase specificity make the policy message clearer, so that it requires less interpretation" (p. 440). The limitation inherent in this view is the assumption that the program to be implemented is optimal in nature at the onset of implementation and that it requires few if any adjustments to accommodate local contexts.

## *2. Include outcome measures that gauge school climate and administrative support.*

At a time when much is known about school and district conditions conducive to the success of innovations, it is naïve to ignore such factors and assume that a program will have the same positive effect on all adopters. Darling-Hammond (2005) has pointed to tensions between top-down reform mandates and the need for autonomous school communities to meet the mandates. These tensions pull “educational systems in contradictory directions” (p. 366), she contends; instead, government policies “should be constructed in ways that maintain the delicate balance between external standards that press for improvement and the school autonomy needed to create an engine for internal change” (p. 375). Too often, evaluations ignore such factors and gather instead the low-hanging fruit of gross achievement impact. Fortunately, elegant approaches such as hierarchical linear modeling and structural equation modeling are now available to identify systemic factors that foster or impair optimal implementation, so long as rich data are available at the level of teacher and school.

## *3. Include multiple designs and aggregate the results.*

We do not view variety in designs and measures as inherently undesirable at scale, but rather as an opportunity to view an initiative through multiple lenses. Meta-analyses have been used to account for variations in instrumentation employed in, for example, Title I (Borman & D’Agostino, 1996), and they hold promise for future initiatives. A caveat, however, is that when indicators of impact are averaged across many sites, the results may well mask valuable lessons that could have been used to improve the program. Reading First has occasioned state-level investigations, usually carried out by teams of university-based researchers, that have varied widely in the questions pursued and the methods employed to answer them. Studies conducted in Michigan (Carlisle, Cortina, & Zeng, 2010), Pennsylvania (Bean, Draper, Turner, & Zigmond, 2010), Florida (Foorman, Petscher, Lefsky, & Toste, 2010), and Utah (Dole, Hosp, Nelson, & Hosp, 2010) give testimony to the rich information that a large-scale initiative can yield—information that could contribute to the improvement of the initiative, or at least ground a more reasoned judgment that it be discontinued.

We suggest that a national evaluation based on a cluster of studies smaller in scale is one key to identifying successful projects and to generating lessons that can be used formatively. Aggregated findings based on meta-analyses and best-evidence syntheses can still be used to provide policy makers with an overall index of program impact. This approach was taken by Borman, Hewes, Overman, and Brown (2003) in their meta-analysis of projects implementing Comprehensive School Reform (CSR) models. The existence of more than a single model required individual consideration of each (in addition to a combined analysis to gauge overall program effectiveness). In a sense, the presence of multiple CSR models is analogous to variations in the manner of Reading First implementation across states, districts, and schools. Although far more varied, these instantiations of the program afford largely unrealized opportunities to look for shared characteristics of the most effective projects. Crawford and Torgesen (2007) have in fact recommended identifying the most

successful projects at the school level and then determining characteristics of these projects that are not present in less successful ones. This approach is tantamount to the process-product research conducted at the teacher level during the 1970s and 1980s, and we suggest that it has excellent potential at the level of the school. This time, though, it could also involve rigorous qualitative case-study approaches that would broaden the scope of evaluation (Hamilton et al., 2003) and embody the preplanned mixed-methods approaches advocated by Johnson and Onwuegbuzie (2004). Grissmer, Subotnik, and Orland (2008) argue persuasively for the inclusion of multiple methods even in randomized experiments. We agree that the use of more than a single lens will provide the best perspectives on not only whether but also how and why a program is working.

## *4. Account for the length of implementation.*

In appraising data from many projects that implemented Comprehensive School Reform (CSR) models, Borman, Hewes, Overman, and Brown (2003) gauged effectiveness not merely in terms of achievement growth but also as a function of how long a model had been in place. They reported that the most promising results were from schools where programs had been implemented for 5 years or more. Although this finding is only one example in a well-documented history of studies examining the impact of reform initiatives in terms of how long they have been in place, length of implementation was not a factor in the Reading First national evaluation. Systematic efforts to account for length of implementation as a potential factor related to the effectiveness of an initiative are admittedly contrary to the demands of policy makers for rapid estimates of impact, but the evidence suggests that these demands should be resisted, or at least mitigated.

## **Conclusion**

Given the dismal history of federal reform initiatives, however well intentioned, it is time to embrace evaluations that are more recursive and elastic. Such evaluations can contribute in a formative manner to the improvement of these initiatives by yielding nuanced results that not only reach but help explain the bottom line. To ignore this option and cling to the all-or-nothing paradigms now in vogue is to deny policy makers the information they need to ground prudent decisions. At this writing, ambitious new federal initiatives in literacy are either in the planning stages or now under way. For example, the K–12 Literacy Education for All, Results for the Nation (LEARN) Act is far more comprehensive than Reading First, as is the Race to the Top initiative, which incorporates more subject areas.

Ensuring that these and other large-scale programs are informed by more useful evaluations is not too much for taxpayers, educators, and children to expect. Such a change, we believe, would acknowledge the incredible contributions of teachers who choose to work under the most challenging conditions and to serve the nation’s poorest children. When their work has been successful, they do not deserve to lose support or face ridicule from opponents because a program has been deemed not to “work.” This is what happened in the case of Reading First, and taking the steps necessary to avoid such an outcome in the future may be its most important lesson.

## REFERENCES

- Allington, R. L. (Ed.). (2002). *Big brother and the national reading curriculum: How ideology trumped evidence*. Portsmouth, NH: Heinemann.
- Bean, R., Draper, J., Turner, G., & Zigmond, N. (2010). Reading First in Pennsylvania: Achievement findings after five years. *Journal of Literacy Research, 42*, 5–26.
- Borman, G. D., & D'Agostino, J. V. (1996). Title I and student achievement: A meta-analysis of federal evaluation results. *Educational Evaluation and Policy Analysis, 18*, 309–326.
- 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.
- Carlisle, J. F., Cortina, K. S., & Zeng, J. (2010). Reading achievement in Reading First schools in Michigan. *Journal of Literacy Research, 42*, 49–70.
- Center on Education Policy. (2006). *Keeping watch on Reading First*. Washington, DC: Author. Retrieved from <http://www.cep-dc.org>
- Center on Education Policy. (2007). *Reading First: Locally appreciated, nationally troubled*. Washington, DC: Author. Retrieved from <http://www.cep-dc.org>
- Crawford, E., & Torgesen, J. (2007). *Teaching all students to read: Practices from Reading First schools with strong intervention outcomes*. Tallahassee: Florida Center for Reading Research. Retrieved from [http://www.justreadflorida.com/reading\\_first.asp](http://www.justreadflorida.com/reading_first.asp)
- Darling-Hammond, L. (2005). Policy and change: Getting beyond bureaucracy. In A. Hargreaves (Ed.), *Extending educational change: International handbook of educational change* (pp. 362–387). Dordrecht, The Netherlands: Springer.
- Darling-Hammond, L. (2006). Securing the right to learn: Policy and practice for powerful teaching and learning. *Educational Researcher, 35*(7), 13–24.
- Desimone, L. (2002). How can Comprehensive School Reform models be successfully implemented? *Review of Educational Research, 72*, 433–479.
- Dole, J. A., Hosp, J. L., Nelson, K. L., & Hosp, M. K. (2010). Second opinions on the Reading First initiative: The view from Utah. *Journal of Literacy Research, 42*, 27–48.
- Duncan, A. (2009, June 15). *Duncan offers stimulus funds for states to develop rigorous assessments linked to common standards* [Press release]. Washington, DC: U.S. Department of Education. Retrieved from <http://www.ed.gov/news/pressreleases/2009/06/06152009a.html>
- Edelsky, C. (1990). Whose agenda is this anyway? A response to McKenna, Robinson, and Miller. *Educational Researcher, 19*(8), 7–11.
- Eisenhart, M., & Towne, L. (2003). Contestation and change in national policy on “scientifically based” education research. *Educational Researcher, 32*(7), 31–38.
- Foorman, B. R., Petscher, Y., Lefsky, E. B., & Toste, J. R. (2010). Reading First in Florida: Five years of improvement. *Journal of Literacy Research, 42*, 71–93.
- Gamse, B. C., Jacob, R. T., Horst, M., Boulay, B., & Unlu, F. (2008). *Reading First impact study: Final report* (NCEE 2009-4038). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences. Retrieved from [http://ies.ed.gov/ncee/pdf/20094038\\_1.pdf](http://ies.ed.gov/ncee/pdf/20094038_1.pdf)
- Grissmer, D. W., Subotnik, R. F., & Orland, M. (2008). *A guide to the use of randomized control trials (RCTs) in assessing intervention effects: The promise of multiple methods*. Retrieved from <http://www.apa.org/ed/schools/cpse/intervention-narrative.pdf>
- Hamilton, L. S., McCaffrey, D. F., Stecher, B. M., Klein, S. P., Robyn, A., & Bugliari, D. (2003). Studying large-scale reforms of instructional practice: An example from mathematics and science. *Educational Evaluation and Policy Analysis, 25*, 1–29.
- Higgins, J. P., Jr. (2007, April, 20). *Testimony of John P. Higgins, Jr., Inspector General, U.S. Department of Education, before the Committee on Education and Labor, U.S. House of Representatives*. Washington, DC: U.S. Department of Education. Retrieved from <http://www.ed.gov/about/offices/list/oig/audit/rpts/stmt042007.doc>
- Institute for Education Sciences. (n.d.). *Reading for Understanding Research Initiative*. Retrieved from <http://ies.ed.gov/ncee/projects/program.asp?ProgID=62>
- International Reading Association. (2006). *Position statement on Reading First*. Newark, DE: Author. Retrieved from <http://www.reading.org/General/AboutIRA/PositionStatements/ReadingFirstPosition.aspx>
- Jackson, R., McCoy, A., Pistorino, C., Wilkinson, A., Burghardt, J., Clark, M., & Schmidt, S. R. (2007). *National evaluation of Early Reading First: Final report*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences. Retrieved from <http://ies.ed.gov/ncee/pubs/20074007/index.asp>
- Johnson, R., & Onwuegbuzie, A. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26.
- Kameenui, E. J., Fuchs, L., Francis, D. J., Good, R., O'Connor, R. E., Simmons, D. C., & Torgesen, J. K. (2006). The adequacy of tools for assessing reading competence: A framework and review. *Educational Researcher, 35*(4), 3–11.
- King, J. A. (2003). Evaluating educational programs and projects in the USA. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (Part 2, pp. 721–732). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- McDaniel, J. E., Sims, C. H., & Miskel, C. G. (2001). The national reading policy arena: Policy actors and perceived influence. *Educational Policy, 15*(1), 93–114.
- Miskel, C., & Song, M. (2004). Passing Reading First: Prominence and processes in an elite policy network. *Educational Evaluation and Policy Analysis, 26*, 89–109.
- National Center for Education Evaluation and Regional Assistance. (n.d.). *NCEE evaluation studies*. Retrieved from <http://ies.ed.gov/ncee/projects/>
- National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.
- Office of Inspector General. (2006). *The Reading First program's grant application process: Final inspection report*. Washington, DC: U.S. Department of Education. Retrieved from <http://www.ed.gov/about/offices/list/oig/aireports/i13f0017.pdf>
- Porter, A. C., Floden, R., Freeman, D., Schmidt, W., & Schille, J. (1988). Content determinants in elementary school mathematics. In D. Grows & T. Cooney (Eds.), *Perspectives on research on effective mathematics teaching* (pp. 96–113). Reston, VA: National Council of Teachers of Mathematics.
- Pressley, M. (2005). Balanced elementary literacy instruction in the United States. In N. Bascia, A. Cumming, A. Datnow, K. Leithwood, & D. Livingstone (Eds.), *International handbook of educational policy* (Part 2, pp. 645–660). Dordrecht, The Netherlands: Springer.
- Reading First Federal Advisory Committee. (2008). *Response to the Reading First impact study interim report*. Washington, DC: U.S. Department of Education. Retrieved from <http://www.ed.gov/programs/readingfirst/statement.pdf>

- Reinking, D., & Bradley, B. A. (2004). Connecting research and practice using formative and design experiments. In N. K. Duke & M. H. Mallette (Eds.), *Literacy research methodologies* (pp. 149–169). New York: Guilford.
- Roller, C. (2009). Public policy and the future of reading comprehension research. In S. E. Israel & G. G. Duffy (Eds.), *Handbook of research on reading comprehension* (pp. 658–667). New York: Routledge.
- Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, 31(7), 15–21.
- Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, 37, 5–14.
- Stein, M., Berends, M., Fuchs, D., McMaster, K., Saenz, L., Yen, L., & Compton, D. L. (2008). Scaling up an early reading program: Relationships among teacher support, fidelity of implementation, and student performance across different sites and years. *Educational Evaluation and Policy Analysis*, 30, 368–388.
- Stern, S. (2008). *Too good to last: The true story of Reading First*. Washington, DC: Thomas B. Fordham Institute.

## AUTHORS

**MICHAEL C. MCKENNA** is Thomas G. Jewell Professor of Reading at the University of Virginia, Department of Curriculum, Instruction and Special Education, Bavaro Hall, 417 Emmet Street South, Charlottesville, VA 22904; [mcm7g@virginia.edu](mailto:mcm7g@virginia.edu). His research focuses on reading attitudes, content area reading, literacy coaching, early reading, and technology applications in literacy.

**SHARON WALPOLE** is an associate professor at the University of Delaware, School of Education, Willard Hall 134C, Newark, DE 19716; [swalpole@udel.edu](mailto:swalpole@udel.edu). Her research focuses on literacy policy, literacy coaching, and early literacy development and assessment.

Manuscript received October 5, 2009

Revision received May 24, 2010

Accepted June 4, 2010