Response to Slavin, Cheung, Groff, and Lake (2008). Effective Reading Programs for Middle and High Schools: A Best-Evidence Synthesis

Thank you for the opportunity to respond to "Effective Reading Programs for Middle and High Schools: A Best-Evidence Synthesis," published in the summer 2008 issue of *Reading Research Quarterly* (Volume 43, Number 3, pp. 290–322).[1] The authors—Robert Slavin, Alan Cheung, Cynthia Groff, and Cynthia Lake—described the purpose of their review as follows: "applying consistent methodological standards" intended "to provide fair comparisons among the achievement effects of the full range of approaches available to educators and policymakers" by placing results of studies on a common scale (p. 291).

Further, they aimed to "provide educators with meaningful, unbiased information that they can use to select programs most likely to make a difference with their students" (p. 292). Although we endorse these goals and admire the scope of their undertaking in this comprehensive review of a wide range of programs and approaches for adolescent literacy development, we also need to be mindful of the unprecedented influence of such reviews in the current context. We need look no further than the unparalleled influence wielded by the National Reading Panel (NRP) review (National Institute of Child Health and Human Development, 2000) of reading research on educational policy and practice to know that in this era, the mantle of "scientifically based" research is increasingly adopted to persuade educators and policymakers as to what works in education—with accompanying pressure to demonstrate adequate yearly progress through accountability and sanctions under No Child Left Behind. Therefore, we believe it is incumbent on the field to scrutinize research syntheses, given their likely influence on research, policy, and practice. We are reminded that the topics comprising the NRP review, determined by panelists' research interests, later wound up as the "five foundations of literacy" mandated by programs under Reading First. We are further reminded that Camilli, Vargas, and Yurecko (2003) conducted a reanalysis of the NRP meta-analysis and determined that tutoring methods and language-rich approaches had as great an effect size as did systematic phonics instruction in early reading achievement, yet these were absent from the NRP conclusions and neglected in later mandates based on the NRP meta-analysis. This history reinforces the need to critically examine reviews and syntheses of this kind.

The intent of our letter is not only to point out some of the strengths of the Slavin et al. synthesis but also to use their work to raise some questions about the logic of the approach and the conclusions drawn in the article and thereby open up a broader discussion in the field. If we understand the "best-evidence synthesis" as an attempt to distill from individual studies something knowable about approaches that work to improve adolescent literacy, then we also need to closely examine possible sources of error to avoid drawing mistaken conclusions. In this case, we are concerned with the way programs were categorized and pooled for analysis and with methodological issues and inconsistencies in the treatment of outcomes for varied learners in varied contexts.

Strengths of the Synthesis

The Slavin et al. review has much to recommend it. Using systematic procedures, the authors identified 33 experimental or quasi-experimental studies relevant to middle and high school reading-improvement programs. Systematic reviews have features that traditional narrative reviews do not. For example, the authors have documented the citations of all of the studies that they excluded from their review and their specific rationale for doing so. Of note in this review, no studies of secondary reading curricula, defined as secondary reading textbooks commonly in use, met the authors' inclusion criteria. Slavin et al. do a service to the field in drawing attention to this fact, which, as they point out, is surprising "given the widespread use of such programs in middle and high schools throughout North America" (p. 294). Table 4 in the article lists all of the programs and curricula for which no qualifying studies were available, and the Appendix lists studies by program that were not included in the review, as well as the reasons for their exclusion. Systematic procedures such as these are important in limiting biases, and the explicitness of systematic reviews permits readers to understand the decision making by the review authors and the potential impact of such decisions.

Systematic procedures are particularly critical in those reviews that include studies with considerable

(continued)

variability of effects. If all studies reported the same outcome, then one study would represent the findings quite well, and systematic procedures to identify, retrieve, and analyze all relevant studies would be less important. But because homogeneity of outcomes is rarely the norm (in fact, even in the same study, there is often a great variation in outcomes), it is even more critical that systematic procedures are undertaken to identify all of the relevant evaluations. As the Slavin et al. review has shown, there is a great variation of impacts across this set of program evaluations, ranging from negative effects (in which students not exposed to treatment do better than do those who receive treatment) to very small effects, and finally, to larger, nontrivial effects in favor of the reading intervention. By being systematic and not selective, Slavin and his colleagues have taken steps toward reducing the kinds of potential biases that could lead reviewers toward one conclusion or another. There are some other methodological approaches used in this review, however, that raise important questions to consider.

Methodological Concerns

Forming Analytic Categories

In a meta-analysis or systematic review of this type, how investigators group studies together is critical. Looking across the landscape of programs for adolescent reading improvement, investigators must decide what the important properties that differentiate and define each category are, as well as have clear and transparent rules for assigning each of the programs they review to the analytically constructed categories. The investigators must therefore know the reading program features well enough to make these categorical decisions confidently and make their assignment criteria adequately transparent to the reader.

Creating broad program categories is a common strategy in reviews of this type. Categorization can be helpful when it simplifies a large amount of data (such as those from 33 studies) into policy- and practice-relevant categories. Generally, meta-analytic practice is to report a global effect size across all programs within that category. But categories can be formed a number of ways, and each of those different constellations can produce a different "bundle" or set of studies and different effect sizes. For example, if you lump one or two studies with very small or negative effects into a single category, particularly when the number of studies in that category is small, it will bring down the overall effect size that is reported. So it is important to look carefully at how program categories are created.

Several examples from this review highlight problems of program categorization. For example, Read 180 and Voyager Passport were categorized by Slavin et al. as mixed methods rather than as computer-assisted instruction programs, despite sharing all of the features of computer-assisted instruction, because they are intended to serve as complete literacy interventions rather than as supplemental programs. Is the primary defining feature of this mixed-method category completeness, then, or an assumption about the instructional time likely devoted to the program, or the absence of other instruction that the computer-assisted instruction is assumed to supplement?

Similarly, the primary defining feature of Slavin et al.'s instructional-process programs is the focus on professional development to change teachers' instructional approaches, yet the authors do not describe the professional-development components of any of the programs categorized as instructional process. Instead, these programs are further subcategorized as (a) cooperative learning, (b) strategy instruction, and (c) comprehensive school reform, based on features of instruction rather than professional development. What defines the category, according to the authors, is therefore not part of the description in the review, yet we know there are significant differences in not only the quality but also the impact of varied methods of professional development (Wei, Darling-Hammond, Andree, Richardson, & Orphanos, 2009). And because Read 180 is described as having a professional development component as well (the program with the longest description of its professional development in this review, p. 295), the reader is left without any understanding as to why it is categorized as mixed methods rather than as instructional process.

Moreover, the three subgroups of cooperative learning, strategy instruction, and comprehensive school reform within Slavin et al.'s instructional-process category raise further questions for us. To illustrate, WestEd's own Reading Apprenticeship Academic Literacy program centers on metacognitive conversation and collaborative meaning making in small group-learning arrangements (Schoenbach, Greenleaf, Cziko, & Hurwitz, 1999). Read 180 also involves collaborative group work. Although we acknowledge the very particular definition Slavin et al. have applied to cooperative learning in their own work, we are left wondering why these programs

were not also categorized with cooperative learning approaches. What defines them, instead, as strategy instruction or as mixed-methods approaches? Conversely, Student Team Reading is categorized as a cooperative learning program, despite its instructional focus, similar to Reading Apprenticeship Academic Literacy, on "explicit teaching of metacognitive strategies" (p. 301). This raised a critical question about the conclusions of the review: Do the reviewers' categories reflect real and meaningful differences between programs and therefore support the conclusions drawn about the relative impact of these different categories of adolescent literacy programs?

Pooling Impacts

One of the difficulties in reviews and meta-analyses, particularly in education, is how to handle multiple cohorts, grades, and outcomes from the same study report. On the one hand, it is recommended that each study should be represented by a single effect size. On the other hand, if a study has three different cohorts (e.g., three consecutive years of entering-student cohorts) and two different grades (e.g., seventh and eighth grade), does combining them obscure important distinctions between cohorts and grades? Although there are no clear statistical or methodological rules for lumping or splitting these effects, being consistent and explicit is good practice (Littell, Corcoran, & Pillai, 2008).

As one of the goals of the Slavin et al. article was to create a common scale or basis for comparisons across programs and instructional categories, the article promotes some healthy investigation into how the development of a common scale was accomplished. Although there is not much detail on how the authors planned to handle multiple outcomes, the results indicate that when multiple effect sizes are reported for multiple grades, multiple cohorts, multiple tests, and multiple reading measures, the authors appear to create effect sizes for each of these and then average them across the entire study for a single program effect. However, this general handling does not hold consistently across all 33 studies. For example, although they included state standardized test scores for other studies, they omitted the AIMS (Arizona's Instrument to Measure Standards) in calculating the mean effect size for studies of Read 180 by White, Haslam, and Hewes (2006) and Johnson, Haslam, and White (2006). For the evaluation of Benchmark Detective, two effect sizes were provided separately for the two cohorts but were not combined (see Table 3, p. 303). We are not sure of the impact of these decisions, but any lack of consistency in how data from individual studies are handled raises concern.

Besides the consistency issue, such pooling of multiple effects raises other questions. The authors averaged the effect sizes for four grades of data that were reported in the Mims, Lowther, Strahl, and Nunnery (2006) study of Read 180. Research has indicated that reading growth typically slows as students mature (e.g., MacMillan, 2000), raising a concern about the impact of combining these effect sizes across so many grade levels. We note that only 5 of the 33 studies reviewed included high school students and that the effect sizes reported for interventions targeting high school students were generally smaller than were those for middle school students. This may be related to the observed asymptotic growth of reading skills. We wonder whether the interpretation of effect sizes should take into account the slowing growth rate for older students on reading outcomes. This also raises a question of interest for the field more generally: Does the construct of adolescent literacy make needed and meaningful distinctions among students and their needs as it is currently applied, referring to students spanning the 6th through 12th grades?

In addition, pooling effect sizes together may also mask important differences in the student samples for these studies. This is a thorny issue for reviews and meta-analyses of all types. Usually such reviews report an effect size to represent the full sample and do not attempt to report effect sizes by important subgroups, such as different types of students. But considerable information is lost in such aggregate comparisons. For example, studies may report outcomes on specific groups of students, such as English-language learners or students with learning disabilities or receiving special education services, and in the Slavin et al. review, the impacts from these studies were averaged with those of students who are defined demographically (e.g., "mostly African American"), geographically (e.g., "rural Jefferson County"), and/or academically (e.g., "remedial"; see Tables 1, 2, and 3). Mims et al. (2006) and Caggiano (2007) both found negative effect sizes for Read 180 with some groups of low-performing African American students, although Caggiano (2007) and Woods (2007) reported sizable positive outcomes for other such groups. Kemple, Herlihy, and Smith

(continued)

(2005) reported a negative effect size for high poverty, mostly African American students in Philadelphia using the Talent Development High School model. Other studies in the review showed a greater benefit for English learners or special education students. We understand that there may be too few studies to break the evidence down by specific student characteristics (i.e., the small-cell problem), yet, it is this kind of evidence that administrators and policymakers will need to make fine-grained decisions about which programs to implement with the populations of students they serve.

We have similar concerns about pooling the effect sizes from different outcome measures. In several cases, studies reported multiple outcomes, for example, reading comprehension and vocabulary, or, in the case of Peer-Assisted Learning Strategies, four subscales of the Woodcock-Johnson III: Letter–Word Identification, Passage Comprehension, Word Attack, and Reading Fluency. Peer-Assisted Learning Strategies, like some of the other programs reviewed, is a multicomponent intervention. For three of the outcome measures, Calhoon (2005) found moderate to large positive effect sizes. These were pooled with the negative effect size found for the Reading Fluency measure to give the study an overall pooled effect size. Similarly, Slavin et al. pooled a positive effect size found by Losh (1991) for comprehension with a negative effect size for vocabulary. For other studies in the review, Slavin et al. combined comprehension and vocabulary outcomes. Yet these varied findings seem to provide important information for making decisions about targeting interventions to particular skill development needed by particular groups of students.

In addition, pooling different outcomes obscures some nuances in the original studies. In the Enhanced Reading Opportunities (ERO) experimental study, which included a test of WestEd's Reading Apprenticeship Academic Literacy as well as the Xtreme Reading program, MDRC and American Institutes for Research identified reading comprehension as the primary outcome (Kemple et al., 2008). Thus, combining vocabulary outcomes (a secondary outcome of the study) with reading comprehension does not reflect the target of the intervention. Averaging the primary and secondary outcomes gives equal weight to both, which was not done in the original evaluation. Research has also indicated that state standardized tests vary tremendously in their quality and sensitivity to various targeted reading proficiencies. As we mentioned earlier, in some cases, state tests were combined with other standardized tests to calculate mean effect sizes for programs. How might the differences in these different outcome measures affect the findings?

Context and Implementation Variables and the Nature of the Counterfactual

Related to our concerns about the ways data were pooled, we wonder whether the review should have included more information and discussion about the contexts in which the various studies took place. Contextual information may help explain why Read 180 produced large positive effect sizes for some students but not others. And such information might alert district decision makers to the factors necessary to allow any particular intervention to succeed. Having been part of the team that provided program implementation support for the ERO study of the Reading Apprenticeship Academic Literacy intervention (Kemple et al., 2008), we know that the first year of the study involved considerable costs to program fidelity, including delays due to teacher assignment and training, student assessment, and student assignment to conditions. The intervention did not begin until mid-November, instead of the beginning of the school year. These problems were reported with the first-year data, and larger effect sizes were reported for sites with higher implementation of program elements. This experience has increased our caution in drawing conclusions about program effects absent information about the nature and degree of program implementation. These are understandably difficult to capture in reviews of this type, but again, raise some caution for us when interpreting findings.

Our reading of the Slavin et al. review highlights the importance of the nature of the control or comparison condition. For example, although it was a large-scale experiment, the ERO study was unique among randomized controlled trials because participation in the program was funded by federal grants that supported the restructuring of high schools into small learning communities. Without comparative information about the conditions under which the other large-scale experimental studies in the review took place, we cannot know if the ERO study is more or less representative of typical school circumstances compared with the others also reviewed in the article.

These unreported contextual and methodological factors become all the more important when trying to make sense of the smaller effect sizes reported for

high school students because in four of the five studies involving high school students, implementation of the adolescent literacy program took place in the context of a large-scale high school reform. However, these four studies, Kemple et al.'s (2008) ERO study of Xtreme Reading and Reading Apprenticeship Academic Literacy and the two studies of Talent Development High School, differed in that the control group for the ERO study was drawn from students within the restructured high schools, whereas the control group for the studies of Talent Development High Schools were drawn from high schools that were not involved in restructuring or reform. In the case of ERO, control students received instruction as usual within the small learning communities—a reform condition—while the treatment group received either the Xtreme Reading or Reading Apprenticeship Academic Literacy interventions. In the Talent Development High School studies, both control and treatment students received a two-period block of literacy instruction, the latter in the context of whole-school reform. These studies might best be characterized as investigations of value added: In the case of ERO, the value added was of supplemental literacy instruction to whole-school reform, while in the Talent Development High School case, the value added consisted of the high school reform model and its particular adolescent literacy approach.

Summary

Perhaps the most important contribution made by the Slavin et al. review is that they have clearly outlined the need for more rigorous studies of reading interventions across different grades and student populations, with attention to the contextual factors that influence implementation and thus program outcomes. It is only through such contextual information that we can begin to really understand the impact of such programs with different types of students at different points in their reading development. In the context of recent federal funding of large-scale, randomized trials, this letter is therefore a call for strong qualitative methods and reporting on the contextual factors that influence outcomes, to enhance the ability of such studies to provide educators with the kind of meaningful, unbiased information that they will need to select programs most likely to make a difference with their students.

The Slavin et al. review also raises important questions for those who are the future producers of such reviews. Although common meta-analytic practice is to report global effect sizes across multiple grades, cohorts, student populations, and outcomes, practitioners and decision makers in the schools are going to be interested in more differentiated effects. This is a challenge in meta-analysis, particularly with smaller numbers of studies, but hopefully the funding of rigorous studies by the Institute of Education Sciences and other agencies will provide more "raw data" for future reviewers to use to drill down more deeply on their review sample. The utility of such reviews to educators, as we have argued, will not depend primarily on the ranking of different programs with varied contexts of study but rather on their explicit treatment of the context as a critical variable in the review.

As with any study or review, the more closely we examined Slavin et al.'s synthesis, the more questions it raised. We realize that with every question and critique, we have made the systematic review of evidence for program effectiveness both more complex and less possible. We are reminded, however, of the infamous First Grade Studies, which attempted to find the approach that worked "best" for beginning readers (Dykstra, 1968). More than 40 years later, we wonder whether we are repeating the folly of looking for one best solution when we know that what educators need is an array of tools and approaches and the information and capacity to choose these tools and approaches based on the particular needs of particular learners and circumstances (Pearson, 1997; Searfoss, 1997). What then can we conclude from a systematic review of this kind, understanding that it is drawing from studies of different programs, used with different cohorts of students, with differing learner needs and characteristics, under very different time frames and circumstances, and comparing the outcomes to different sorts of counterfactuals? Perhaps the one kernel of truth that emerges from the review of these 33 studies is that in almost all cases, doing something to build literacy proficiencies for students in middle and high schools turns out to be better than doing nothing. That is a message our secondary schools need to hear.

Cynthia Greenleaf, WestEd, San Francisco, California, USA.

Anthony Petrosino, WestEd, Regional Educational Laboratory Northeast and Islands, San Francisco, California, USA.

(continued)

Notes

[1] We applaud Robert Slavin for acknowledging possible conflicts of interest in being part of the development team for several of the programs reviewed in this article. We follow his example by acknowledging that we work for WestEd, a research, development, and service agency that developed one of the programs, Reading Apprenticeship, that was reviewed in this same article.

References

Caggiano, J.A. (2007). *Addressing the learning needs of struggling adolescent readers: The impact of a reading intervention program on students in a middle school setting*. Unpublished doctoral dissertation, The College of William and Mary, Williamsburg, VA.

Calhoon, M.B. (2005). Effects of a peer-mediated phonological skill and reading comprehension program on reading skill acquisition for middle school students with reading disabilities. *Journal of Learning Disabilities*, *38*(5), 424–433. doi:10.1177/00222194050380050501

Camilli, G., Vargas, S., & Yurecko, M. (2003, May 8). Teaching children to read: The fragile link between science and federal education policy. *Education Policy Analysis Archives*, *11*(15). Retrieved March 20, 2007, from epaa.asu.edu/epaa/v11n15/

Dykstra, R. (1968). Summary of the second-grade phase of the Cooperative Research Program in primary reading instruction. *Reading Research Quarterly*, *4*(1), 49–70. doi:10.2307/747097

Johnson, J., Haslam, M., & White, R. (2006). *Improving student literacy in the Phoenix Union High School District, 2005–06*. Washington, DC: Policy Studies Associates.

Kemple, J.J., Corrin, W., Nelson, E., Salinger, T., Herrmann, S., Drummond, K., et al. (2008, January). *The enhanced reading opportunities study: Early impact and implementation findings* (NCEE 2008-4015). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Kemple, J.J., Herlihy, C.M., & Smith, T.J. (2005, May). *Making progress toward graduation: Evidence from the Talent Development High School model*. New York: MDRC. Retrieved January 11, 2009, from www.mdrc.org/publications/408/overview.html

Littell, J.H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. Oxford, England: Oxford University Press.

Losh, M.A. (1991). *The effect of the strategies intervention model on the academic achievement of junior high learning-disabled students*. Unpublished doctoral dissertation, University of Nebraska–Lincoln.

MacMillan, P. (2000). Simultaneous measurement of reading growth, gender, and relative-age effects: Many-faceted Rasch applied to CBM reading scores. *Journal of Applied Measurement*, *1*(4), 393–408.

Mims, C., Lowther, D., Strahl, J.D., & Nunnery, J. (2006). *Little Rock School District READ 180 evaluation: Technical report*. Memphis, TN: The University of Memphis, Center for Research in Educational Policy.

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.

Pearson, P.D. (1997). The first-grade studies: A personal reflection. *Reading Research Quarterly*, *32*(4), 428–432. doi:10.1598/RRQ.32.4.5

Schoenbach, R., Greenleaf, C., Cziko, C., & Hurwitz, L. (1999). *Reading for understanding: A guide to improving reading in middle and high school classrooms*. San Francisco: Jossey-Bass.

Searfoss, L.W. (1997). Connecting the past with the present: The legacy and spirit of the first-grade studies. *Reading Research Quarterly*, *32*(4), 433–438. doi:10.1598/RRQ.32.4.6

Wei, R.C., Darling-Hammond, L., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the United States and abroad*. Dallas, TX: National Staff Development Council.

White, R.N., Haslam, M.B., & Hewes, G.M. (2006, July). *Improving student literacy in the Phoenix Union High School District 2003–04 and 2004–05: Final Report*. Washington, DC: Policy Studies Associates.

Woods, D.E. (2007). *An investigation of the effects of a middle school reading intervention on school dropout rates*. Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA.