MICHAEL R. VITALE, East Carolina University;
THEODORE S. KANIUKA, Fayetteville (NC) State University

Adapting a Multiple-Baseline Design Rationale for Evaluating Instructional Interventions: Implications for the Adoption of Direct Instruction Reading Curricula for Evidence-Based Reform

Abstract: Present national methodological standards for evaluating the credibility of the design of individual research studies have resulted in findings reporting the pre-post effectiveness of Direct Instruction programs to be eliminated from consideration by educational leaders involved in making curricular decisions intended to advance local school reform initiatives. Presented is an evaluative framework based on the logic of multiple-baseline designs developed in behavior analysis (Sidman, 1960) that, in combination with multi-level statistical analysis, provides a sound basis for the evaluation of the achievement impact of Direct Instruction programs useful to educational leaders faced with making decisions to adopt school curricula to improve student achievement. Summarized is (a) the history and status of Direct Instruction programs as effective tools in school reform, (b) a methodological rationale for the use of a multiple-baseline design framework in the evaluation of Direct Instruction programs, (c) an illustrative application of the proposed methodology for combining existing achievement data to evaluate the impact of Direct Instruction reading programs previously implemented across multiple school sites, and (d) the implications for using the methodological framework in conjunction with other evaluative approaches in support of advocacy for Direct Instruction programs.

A fair statement is that systemic school reform initiatives in the U.S. have been less than successful. Focusing on the extremes of the K-12 reading continuum, the recent National Assessment of Educational Progress (2009) shows no increase in student achievement at the high school level from 1992 to 2009, and the preliminary (Year 2) Reading First component of No Child Left Behind has been evaluated as noneffective (Gamse, Bloom, Kemple, & Jacob, 2008). Moreover, despite continuing to report a high degree of success on reading as assessed through state-developed tests in grades 3-8, many low-achieving states (e.g., Florida, North Carolina) have ignored the substantial declines in the achievement of school-dependent (e.g., low SES) students at the high school level and the related evidence that the state-established proficiency levels set at the upper elementary grades are far too low to be indicators of future student academic success in high school (see Dolan, 2005). Rather than pursuing systemic changes in reform, the instructional initiatives and other reform strategies offered to schools consist of "more of the same" approaches in the form of traditional instructional alternatives, coaching initiatives to support present practices, and

leadership strategies advocated by established and "trusted" sources that guide the reform process for practitioners (e.g., USDOE/ Institute of Education Sciences [IES] What Works Clearinghouse, 2006, 2010d).

Given the present status of school reform, it is not surprising that both research methodologists and educational leaders have raised concerns about evaluating school reforms. Linn (2000) questioned the validity associated with the use of state-developed accountability tests ,and Dolan (2005) found state-established student proficiency levels in reading at the upper elementary levels substantially over predicted subsequent student achievement at the high school levels (i.e., proficiency at the upper elementary level did not imply proficiency at the high school level). In related concerns, Hess (2008) described the problems associated with the simplistic use of accountability data by schools. Ronka, Lachat, Slaughter, and Meltzer (2008) and Reeves (2008) emphasized the need for schools to ask and answer questions relevant to systemic school reform rather than limiting themselves to questions about accountability data provided by state tests (see also Popham, 2008).

Among the complementary reform issues relating to the evaluative use of accountability data (see also Linn, 2005) and test validity (e.g., Barton & Coley, 2008; Brennan, 2004) is the identification of evidence-based interventions which schools are encouraged to use (Slavin, 2008a, 2008b). For example, an important methodological emphasis in recent years has been the use of randomized field trials and multilevel statistical modeling (e.g. Raudenbush & Byrk, 2002) as standards for evaluating the effectiveness of instructional interventions (see USDOE/IES Research Standards, IES, 2010a). In general, there is no argument against the logic of randomized field trial designs and the use of advanced multi-level statistical modeling per se. However, because school reform necessarily is an evolutionary process, the emphasis on randomized field trials as the means for establishing evaluative conclusions regarding the generalization of the effectiveness of instructional interventions artificially limits the decision making of school leaders about the potential value of different instructional interventions (e.g., Burch, 2007; Slavin, 2008a, 2008b; Sloane, 2008). Certainly, for cases in which replicated studies using randomized field trials for a specific intervention exist, sound estimates of the effects of instructional interventions can be obtained. But, at the present time, the availability of such findings is very limited and, by implication, excludes from consideration a wide range of intervention research with potential for the systemic improvement of education outcomes (see IES What Works Clearinghouse, 2010b). In addition, recent developments in quasi-experimental research design question the use of randomized field trials as the only means for establishing causality in circumstances in which randomization is impractical (see Hawkins, Sanson-Fisher, Shaeshaft, D'Este, & Green, 2007; Murnane & Willett, 2011).

We propose an evaluative framework based on traditional multiple-baseline designs developed in the field of behavior analysis (Sidman, 1960) that, in combination with multi-level statistical analysis, provide a basis for the evaluation of Direct Instruction programs. First, we briefly summarize the history and present status of Direct Instruction programs as effective tools in school reform. Next, we provide a methodological rationale for the use of the multiple-baseline design framework in the evaluation of Direct Instruction programs. Third, we present an illustrative application of the design methodology for evaluating the achievement impact of Direct Instruction reading programs implemented across multiple sites. Last, we discuss implications for using the methodological framework in conjunction

with other evaluative approaches to support advocacy for Direct Instruction programs.

Direct Instruction Programs and School Reform

As used here, Direct Instruction (DI) refers to three interdependent components, all associated with the work of researchers initiated with Project Follow Through in the 1960s (see the Association for Direct Instruction [http://www.adihome.org/] and the National Institute for Direct Instruction [http://www.nifdi.org/15/]). The first component consists of a general instructional design model (Engelmann & Carnine, 1991) used to develop and validate instructional curricula. The second component includes the family of instructional curricula developed and validated using the design model (see Grossen, 1988, n.d.; Marchand-Martella, Slocum, & Martella, 2003). The third component is an empirical research base of literature providing evidence for the effectiveness of the principles underlying the instructional design model (see Adams & Engelmann, 1996; Engelmann & Carnine, 1991). At present, many educators are familiar with the grade K-5 *Reading Mastery* program (which evolved from the original grade K-3 DISTAR Reading Series) because it has been widely recognized as a research-validated developmental reading program (e.g., American Federation of Teachers 1999; American Institutes for Research, 2006; Adams & Engelmann, 1996). However, most educators are not familiar with the rich conceptual content in advanced levels of Direct Instruction programs in the different curricular domains.

There is abundant evidence supporting the effectiveness of Direct Instruction curricula. For example, a careful analysis of the evaluation results for Project Follow Through (see Grossen, 1995-96), originally reported by Abt Associates (Stebbins, St. Pierre, Proper, Anderson, & Cerva, 1977), showed the Direct Instruction intervention in grades K-3 composed of DISTAR Reading/Spelling, Language, and Mathematics was effective in raising the academic proficiency of at-risk students to prepare them for future success in grades 4-5 (see Adams, 1995-96; Becker & Engelmann, 1995-96; Gersten, Darch, & Gleason, 1988; Gersten & Keating, 1987; Gersten, Keating, & Becker, 1988; Meyer, 1984). Complementing the Project Follow Through results, Adams and Engelmann (1996) reported a rigorous meta-analysis of research studies involving a wide variety of participants ranging from kindergarten through college-educated adults that documented the effectiveness of applying Direct Instruction across a wide range of instructional content. More recently, Marchand-Martella et al. (2003) summarized evaluative research findings encompassing Direct Instruction curricula across a variety of content areas (e.g., reading, language, writing, thinking, science).

Given the established findings in support of the effectiveness of Direct Instruction programs in general, and the cumulative evaluative findings initiated with Project Follow Through, a possible expectation may be that the use of Direct Instruction programs has been a major focus of school reform initiatives. However, this has not been the case. Despite established evidence of effectiveness, implementation capacity (see Engelmann & Engelmann, 2004), and increased cumulative use with at-risk students (see National Institute for Direct Instruction, 2010a, 2010b), Direct Instruction—for the most part—has been excluded from the school reform movement. As Engelmann (2008) detailed, what evolved into a systemic avoidance of using Direct Instruction programs that began with Project Follow Through has been detrimental to both school reform and the potential achievement of at-risk students.

A number of different perspectives have been offered to address the exclusion of Direct

Instruction from school reform. In a chapter reviewing the effectiveness of Direct Instruction reading, Ellis (2005, Chapter 10) discussed the lack of interest by educators engaged in school reform toward instructional approaches that work versus their preference for approaches with no established validity. From a broader perspective, Kuhn (1966) pointed out how established disciplines were resistant to views threatening possible paradigmatic changes to the widely accepted intellectual representations of the discipline. Consistent with Kuhn's analysis, Hirsch (1999) outlined tactics of disciplinary thinking in education (i.e., the "education thoughtworld") that discouraged practitioners from deviating from established views, for example, the process of artificial characterization (*Direct Instruction consists of kill and drill*) and bipolarization of perspectives (*since kill and drill is bad, Direct Instruction is bad*) to maintain accepted thinking. At a more specific level, Tarver (2004) addressed common "myths" used by educators to discourage the pursuit of gaining an understanding of Direct Instruction programs.

In addition, Engelmann (2008) noted the IES What Works Clearinghouse (2010) purports to list effective instructional programs for practitioners and arbitrarily set 1985 as the earliest year research addressing instructional effectiveness would be included in their present database. This decision eliminated the evaluative evidence from Project Follow Through for the Direct Instruction developmental programs in reading, spelling, language, and mathematics from present school reform initiatives. Despite the emphasis in school reform to only use approaches having empirically-established, research-based validity, Engelmann (2003) and Engelmann, Batemann, and Lloyd (2007) described how the concept of evidence-based validity, as applied in school reform, allows instructional interventions to be justified in terms of effectiveness when they only incorporate fragmented practices associated with published research studies. As a result of allowing such a substitution, instructional interventions are able to avoid the requirement to unambiguously demonstrate effectiveness in engendering student academic achievement outcomes.

A Rationale for Applying Multiple-Baseline Design Frameworks to the Validation of Direct Instruction Programs

One problem implied in the preceding sections was the randomized field trial methodology intended to ensure valid conclusions from individual studies. This is overly restrictive in providing information of value to educational leaders faced with the selection of instructional alternatives in school reform (see also Henig, 2008; Hess, 2008; Ronka et al., 2008). Following an instructional systems perspective (Dick, Cary, & Cary, 2007), educational leaders considering interventions for adoption need to know (a) what performance outcomes are expected to result from an intervention implemented with fidelity and (b) whether the implementation of the intervention is feasible. Although well-conducted, large-scale, randomized field trials could provide such achievement expectation estimates, subsequent replication of such studies is still a necessary requirement for confirming or disconfirming research conclusions. With this in mind, an important question is what sort of an informational framework could educational leaders who are making curricular decisions use in the absence of evaluative findings based on randomized field trials?

Emphasizing the Concept of Replicability in Evaluative Research

The overarching evaluative concept educational practitioners should hold is that replic-

ability of findings is the most important scientific standard for research findings to meet. That is, replicability of findings is the most useful form of evidence-based information of intervention effectiveness, not the findings of any single study, no matter how well such studies are designed. In emphasizing replicability, the logical structure of multiple-baseline designs (see Sidman, 1960) is a far more appropriate design framework for the evaluation of the effectiveness of instructional interventions than traditional group designs because they involve intrastudy replication of the effects of experimental interventions across what Campbell and Stanley (1963) call "time series."

As an example, in a multiple-baseline design involving three experimental units (e.g., schools), baseline data by grade would be obtained for a series of intervals (e.g., tests by years), after which an experimental intervention would be introduced to one randomly-selected unit while data collection continued for the remaining two. Then, after the experimental effect of the intervention on the first selected unit stabilized, the intervention would continue with the first unit and be implemented with a second randomly selected unit. Again data collection would continue until the effect of the second implementation of treatment stabilized. Finally, the intervention would be continued in the first two units and implemented in the third unit. Once the effect in all three units was stabilized, then determining the experimental effect of the treatment would consist of comparing the performance of the three units under baseline (no treatment) conditions with that of the three units after treatment was implemented.

In multiple-baseline designs, the point is to demonstrate that the experimental effect observed after the intervention was implemented (in comparison to the performance obtained under baseline without the treatment being implemented) supports the conclusion that the experimental effect was produced by the instructional treatment. Of course, if the expected effect does not occur, then the intervention is classified as ineffective. From a design standpoint, the important points are that (a) the emphasis in the overall design is on an intrastudy replication of effects and (b) the scope of the design logically subsumes that of randomized field trials in which the assignment of treatment is not time-lagged (i.e., in randomized field trials all treatment interventions are ordinarily implemented at the same point in time).

While logically powerful, multiple-baseline designs typically are not applicable as true experiments in school evaluation settings because of the extensive resource demands needed for multischool, multiyear research projects. However, multiple-baseline design logic is highly applicable as a quasiexperimental design alternative that uses existing data to evaluate the effect of time-lagged, multischool instructional interventions. Given this distinction, multiple-baseline design scenarios fall into two categories. In the first, prior existing baseline data are obtained for all experimental units after which the remainder of the design involving randomized time-lagged introduction of treatment is followed. In the second, illustrated here, all of the existing data, both prior to and after an intervention, are collected from different sites across which an intervention was time-lagged.

The first case clearly represents experimental studies, while the second, correlational studies. The point is, because of their emphasis on replication, both are capable of providing an informational framework of high value to school leaders involved in making curricular decisions in "real time" regarding the potential effectiveness of an instructional intervention.

Using the Logic of Multiple-Baseline Designs as a Model for Evaluating the Effectiveness of Instructional Programs

The inferential conclusion resulting from a successful multiple-baseline design is always in the form of a specified effect (i.e., outcome) that can be produced by implementing a particular intervention. This type of causal inference is consistent across all domains of scientific investigations that utilize experimental research methodology. The reasoning justifying such an inferential conclusion being based on successful and multiple replications is straightforward. In other words, if multiple applications of an intervention have consistently produced a desired effect in the past, then that intervention would be expected to produce a similar effect in comparable settings if implemented with fidelity. Although interpreted probabilistically rather than as a proposition that is true or false, such reasoning is fundamental to the establishment of all scientific knowledge.

Given the preceding rationale, the argument presented here is as follows. Although all experimental studies require implementation of an intervention under the control of researchers, a pattern of pre-existing data following the logic of a multiple-baseline design can yield meaningful inferential conclusions. But to do so, existing data must be scaled to allow integration of the effects of an intervention across different (and independent) settings, and the interventions must be time staggered. Such a rationale provides a means to combine independent research components that, without such integration, would not yield interpretable data. In summary, the argument advanced is: *If an intervention can be shown to produce results in a pre-post intervention context across many independent sites, then, from an evaluative standpoint, the resulting data can be interpreted as providing "causal" evidence of effectiveness, subject to probabilistic qualifications.* From the standpoint of educational leaders considering Direct Instruction programs for possible adoption, such conclusions are sufficient to meet the requirement that an intervention is evidence based.

An Illustrative Example Applying a Multiple-Baseline Design Framework to Reading Achievement Data from Direct Instruction Sites

This section presents an illustrative example of using a multiple-baseline framework that incorporates existing data as *interschool* replications to evaluate the effectiveness of Direct Instruction reading programs. In doing so, existing pre- and post-intervention performance trends from Direct Instruction reading implementations that were reported in summary form on the 2008 Science Research Associates (SRAonline.com) website were analyzed. While the analysis conducted recognizes such findings raise questions regarding sampling bias (i.e., only positive findings are reported), the reported findings within schools represent a form of multiple-baseline design logic amenable to multilevel statistical analysis (i.e., HLM). From the standpoint of replicability, the HLM analysis of the multilevel set of pre-post data provided a means for evaluating the effect of the time-lagged Direct Instruction reading interventions across multiple school sites.

Participants

Participants were 73 elementary and middle schools implementing Direct Instruction reading programs in the K-8 grade range as part of a school reform or achievement improvement initiative during a 13-year time span (i.e., 1996-2008). Each participating school reported multiyear pre- and post-intervention student reading performance by grade and by year as

measured by state-administered or nationally-normed reading tests. The participating schools were demographically diverse (e.g., percentage minority ranged from 3% - 100%, free/reduced lunch ranged from 0% - 99%).

Intervention

The intervention consisted of several Direct Instruction reading programs. Primarily, *Reading Mastery* was implemented in grades K-5, while *Corrective Reading* was implemented in grades 3-8. *Horizons*, a program for grades K-3 in combination with *Corrective Reading* was used in several schools.

Research Design

The researchers considered the set of multi-year data (by grade) from each school as an individual unit. Overall, in this quasiexperimental (i.e., correlational) study, the data-by-year always included prior- and post-intervention results, and the implementation of the interventions across different schools was time distributed. Therefore, the data/design structure of the study fit the logic of a multiple-baseline framework over the 13-year time interval.

Data Collection and Analysis

The achievement data for each school were obtained from descriptive studies (e.g., mean student achievement by grade and by year, prior to- and after-interventions) reported on the 2008 SRA web site. Because the data by school were only descriptive summaries, intraschool error estimates were not available and typical meta-analytic HLM approaches (Raudenbush & Byrk, 2002, Chapter 7) could not be used. As a substitute, within each school (or within grades within schools), achievement data were scaled across years to have means of zero and standard deviations of unity. This insured pre-post implementation trends and postimplementation trends obtained from different achievement measures (e.g., different state or norm-referenced reading tests) across different schools were comparable.

In addition to grade level within school, statistical design/dummy variables (i.e., 0 = preintervention, 1 = postintervention) were used to represent prepost and postimplementation trends. Complementing these data, the percentage of minority students and percentage of students on free/reduced lunch were coded at the school level. Therefore, in the HLM analysis, the pre-post and post achievement trends by school-grade by-year were coded at level 1, while school demographic characteristics were coded at level 2. This multilevel analysis approach allowed each school to serve as its own control with regard to time-lagged, pre- vs. post-intervention achievement differences while the overall analysis addressed the general question of whether the interventions resulted in improved achievement after the intervention was introduced. The multilevel statistical analysis was conducted using HLM6 (Raudenbush, Bryk, & Congdon, 2004).

Results

HLM results found the pre-post difference in standardized student reading achievement was statistically significant, $t\ (573) = 4.04, p < .001$. The percentage of minority students in school, $t\ (573) = -2.34, p < .020$, also was significant, with nonminority students outperforming minority students. The percentage of students receiving free/reduced lunch was not significant. Adding an additional variable into the HLM model representing year of implementation (within school) also showed an accelerated positive achievement trend after the initial year of program implementation ($t\ (570) = 2.30, p < .022$).

The overall results showed the implementation of Direct Instruction reading programs across different school settings engendered a significant improvement of student achievement that was further accelerated after the initial intervention year. Consistent with requirements of a multiple-baseline design, the interventions had been implemented

across different schools in a time-lagged sequence. And, because schools represented a wide variety of geographical and demographic settings, the HLM findings provided strong evidence of replicability of the positive effect of the Direct Instruction intervention. Even though the data may reflect a bias that excluded schools whose Direct Instruction interventions were not successful, the overall point is that the effectiveness of Direct Instruction programs was demonstrated across a substantial number (N= 73) of elementary schools over the 13-year time period.

Discussion and Implications

According to the USDOE/IES, the intent of the What Works Clearinghouse is to provide education practitioners with "*a central and trusted source of scientific evidence for what works in education*" (IES, 2010b, para. 2). The evidence presented by the Clearinghouse for Direct Instruction programs in reading (e.g., *Reading Mastery, Corrective Reading, Horizons*) categorizes them as "*minimally effective*" at best, based on the few studies that met Clearinghouse Standards. In contrast, Stockard (2008) found the Clearinghouse rejected close to 100 studies as being methodologically flawed despite being cited in well-established literature reviews (see also Stockard, 2010). Considering Stockard's comprehensive analysis in conjunction with the Clearinghouse's elimination of Project Follow Through, and other research prior to 1985, Stockard's recommendation that educators use extreme caution in using the evaluations reported by the Clearinghouse is understandable. Coupling the preceding deficiencies with a general lack of response by the Clearinghouse to detailed methodological and policy issues (see Engelmann, 2008) raises significant questions about whether the operational practices of the Clearinghouse are consistent with its stated goal.

At present, in order for educational practitioners to make informed decisions regarding the adoption of Direct Instruction programs for reform initiatives, they must be able to integrate the effectiveness information presented by the Clearinghouse with the extensive body of evaluative research studies reporting the effectiveness of school-based Direct Instruction. However, as an additional perspective, the statistical analysis of interschool patterns of student achievement growth within a multiple-baseline design framework may provide educational practitioners with an alternative for evaluating the effectiveness of Direct Instruction. Such an alternative working framework is not only conceptually straightforward but also consistent with informal strategies educators use to obtain information of program effectiveness from colleagues at other school sites who are using an instructional program of interest (see Burkhardt & Schonfeld, 2003; Fusarelli, 2008; Kochanek & Clifford, 2011).

While the multiple-baseline design presented is not intended to replace either Clearinghouse or individual empirical studies of Direct Instruction, the explicit emphasis of multiple-baseline designs on replicability makes them highly useful to educational leaders in curricular decision-making roles. In this regard, the systematic collection of pre- and post-implementation achievement data on a continuing basis and the reapplication of multi-level analysis of the accumulating achievement patterns would be an increasingly valuable form of evaluative information for educational practitioners. Required for such initiatives would be the establishment of an upgradable database that would accumulate both patterns of achievement data and the dissemination of the associated trends resulting from multilevel analyses as new achievement data are added. In the case of Direct Instruction programs, there is an increasing amount of data potentially available from K-8 schools across a variety of curriculum areas (e.g., reading, writing, spelling, mathematics). Within this analytic context, as positive achievement

trends for with Direct Instruction accumulating from school-based implementations are added to an evaluative database, the result could be an increasingly stronger advocacy for the adoption of Direct Instruction programs by education practitioners.

From a methodological point of view, rather than research initiatives being limited to meeting the requirements for randomized field trials within large-scale studies, the multiple-baseline framework presented here would collect interschool achievement data from time-lagged interventions in a manner that emphasizes replication in a form directly useable by school practitioners. With this in mind, it is important to recognize that for the discipline of education to benefit systemically from the large body of work implementing Direct Instruction programs in schools, researchers must work collaboratively with education practitioners. The cumulative growth of student achievement resulting from multiyear use of Direct Instruction Programs should be documented in a manner that the resulting achievement patterns can be archived, analyzed statistically, and disseminated to practitioners. In using such evaluative data to enhance educator support for Direct Instruction programs, the following perspectives and actions should be considered by school reform leaders considering the adoption of Direct Instruction programs (see Vitale, Romance, & Kaniuka, 2011):

1. *Recognizing that the overall achievement impact of Direct Instruction programs is cumulative.* That is, the achievement impact of Direct Instruction programs improves as a function of increased teacher experience and, even more importantly, the number of years students are in the programs. This perspective implies that the effects of Direct Instruction programs should always be interpreted as dynamically evolving "proof-of-concept" demonstrations rather than as absolute evidence-based conclusions.

2. *Obtaining additional information regarding the intervention in school sites whose demographics are similar to the one for which the intervention is being considered.* In this regard, the type of multiple-baseline study encompassing achievement data from other school sites presented here would provide important preliminary evaluative information. In considering evaluative information in making a decision about whether to adopt a Direct Instruction program, such information should be obtained from demographically similar schools that had used the Direct Instruction curriculum being considered. This information should include the specific outcome measures used, the training and support required for implementation, and an in-depth knowledge of the curricular content students would experience through the program. Essentially, this consists of developing an understanding of key aspects of the intervention based on the experience of those who have used it successfully.

3. *Finalizing the decision to adopt a Direct Instruction program, educational decisionmakers (with appropriate technical assistance) should use available information to estimate the expected achievement outcomes that would result from adoption of the intervention.* In fact, this can be accomplished by substituting the local demographic characteristics (e.g., percentage of minority students) into the HLM model produced. Then, based on such expectations, decisionmakers should insist on the establishment of a strong local evaluative model that would relate achievement outcomes to program level (i.e., lessons completed) along with cumulative grade-by-grade comparisons with achievement from demographically similar schools across years.

4. *Obtaining multiyear evaluative results, comparisons of achievement of schools using Direct Instruction interventions with achievement expectations and with that of comparison schools should*

be reported. If successful, these results also should be submitted to sites that publish effectiveness findings resulting from well-designed evaluative studies.

5. *Developing the management capacity for supporting expansion (scale up) of implementation so that, within the context of a school system, randomized field trials following a multiple-baseline experimental design could be implemented.* In doing so, it is likely that program adoption would follow a time-lagged sequence that would impose an experimental multiple-baseline design involving successive groups of schools over a series of years that would incorporate accepted randomized field trial methodology.

A critical foundation of the preceding argument is that knowledge accrues through the replication of well-designed research or, in the case of education, evaluative studies. Certainly it is important to establish and advocate strong standards such as those of randomized field trials for establishing the credibility of research findings. But it is equally important for educational researchers to work to develop strategies for the credible use of available information by educational leaders involved in the process of making decisions about which interventions to adopt rather than simply excluding extensive patterns of evaluative data because individual studies do not meet idealized requirements of one methodological approach (e.g., What Works in Education, 2010). The application of the multiple-baseline framework presented describes and illustrates one such alternative methodological approach that could provide increasingly powerful decisionmaking support to practitioners without precluding the concurrent use of idealized forms of research design such as randomized field trials presently advocated as a "gold standard" for providing evidence of educational effectiveness.

References

Adams, G. (1995-96). Project Follow Through and beyond. *Effective School Practices*, 15. Retrieved from http://darkwing.uoregon.edu/~adiep/ft/adams.htm

Adams, G., & Engelmann, S. (1996). *Research on direct instruction: 25 years beyond DISTAR.* Seattle, WA: Educational Achievement Systems.

American Federation of Teachers (AFT). (1999). Building on the best, learning from what works: Five promising remedial reading intervention programs. Washington, DC.

American Institutes for Research (2006). *CRAQ Center report on elementary school comprehensive school reform models.* Comprehensive School Reform Quality Center. Washington DC.

Barton, P. E., & Coley, R. (2008). Measuring the achievement elephant. *Educational Leadership, 66*(4), 30-34.

Becker, W. C., & Engelmann, S. (1995-96). Sponsor findings from Project Follow Through. *Effective School Practices,* 15. Retrieved from http://darkwing.uoregon.edu/~adiep/ft/becker.htm

Brennan, R. L. (2004). *Revolutions and evolutions in current educational testing.* Occasional Research Paper #7, Des Moines, IA: The Iowa Academy of Education.

Burch, P. (2007). Educational policy and practice from the perspective of institutional theory: Crafting a wider lens. *Educational Researcher, 36,* 84-95. doi:10.3102/0013189X07299792

Burkhardt, H., & Schonfeld, A. H. (2003). Improving educational research: Toward a more useful, more influential, and better-funded enterprise. *Educational Researcher, 32*(9), 3–14.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago, IL: Rand McNally.

Dick, W. O., Cary, L., & Cary, J. O. (2007). *Systematic design of instruction.* New York: Pearson

Dolan, M. F. (2005). Assessment success today or learning success tomorrow? How a longitudinal perspective helps standards-based accountability systems eliminate the persistent gap between nominal and actual achievement for high school graduates. *Dissertation Abstracts International, 66, 567.*

Ellis, A. (2005). *Research on educational innovations.* (4th ed.). Larchmont, NY: Eye on Education.

Engelmann, S. (2003). The Dalmatian and its spots. *Education Week, 23*(20), 48,34,35.

Engelmann, S. E., & Engelmann, K. F. (2004). Impediments to scaling up effective comprehension school reform models. In Glennan, T. K. (Ed.). *Expanding the reach of education reforms: Perspectives from leaders in the scale-up of educational interventions*

(pp. 107-135). Santa Monica, CA: Rand Corporation.

Engelmann, S. (2008). *Machinations of the What Works Clearinghouse*. Retrieved from http://www.zigsite.com/PDFs/MachinationsWWC%28V4%29.pdf

Engelmann, S., Bateman, B., & Lloyd, J. W. (2007). *Educational logic and illogic*. Eugene, OR: Association for Direct Instruction.

Engelmann, S., & Carnine, D. (1991). *Theory of instruction*. (Rev. ed.). Eugene, OR: Association for Direct Instruction.

Fusarelli, L. (2008). Flying (partially) blind: School leaders' use of research in decision-making. *Phi Delta Kappan 89*, 365–368.

Gamse, B. C., Bloom, H. S., Kemple, J. J., & Jacob, T. J. (2008). *Reading First impact study*. USDOE/Institute of Education Sciences, Washington, DC.

Gersten, R., Darch, C., & Gleason, M. (1988). Effectiveness of a Direct Instruction academic kindergarten for low-income students. *The Elementary School Journal*, 89, 226-240. doi:10.1086/461575

Gersten, R., & Keating, T. (1987). Long-term benefits from Direct Instruction. *Educational Leadership*, *44*(6), 28-31.

Gersten, R., Keating, T., & Becker, W. (1988). The continued impact of the Direct Instruction Model: Longitudinal studies of Follow Through students. *Education and Treatment of Children*, *11*, 318-327.

Grossen, B. (1995-96). The story behind Follow Through. *Effective School Practices, 15*. Retrieved from http://darkwing.uoregon.edu/~adiep/ft/grossen.htm

Grossen, B. (1988). *The research base for Corrective Reading (SRA)*. Columbus, OH: McGraw-Hill.

Grossen, B. (n.d.). *The research base for Reading Mastery, SRA*. New York, NY: Macmillan-McGraw-Hill. Retrieved from http://uncw.edu/wha/hillcrest/documents/THERESEARCHBASEFORREADINGMASTERYbyBonnieGrossen.pdf

Hawkins, N., Sanson-Fisher, R., Shaeshaft, A., D'Este, C., & Green, L. (2007). The multiple baseline design of evaluating population-based research. *American Journal of Preventative Medicine, 33*, 162-168. doi:10.1016/j.amepre.2007.03.020

Henig, J. R. (2008). The spectrum of educational research. *Educational Leadership, 66*(4), 6-11.

Hess, F. M. (2008) The new stupid. *Educational Leadership, 66*(4), 12-17

Hirsch, E. D. (1999). *The schools we need and why we don't have them*. New York, NY: Random House.

Institute of Education Sciences (2006). *WWC evidence review protocol for beginning reading instruction*. USDOE/IES, Washington, DC. Retrieved from http://ies.ed.gov/ncee/wwc/references/idocviewer/doc.aspx?docid=27&tocid=1

Institute of Education Sciences (2010a). *Education Research Grants. CFDA Number: 84.305A*. USDOE/IES, Washington, DC. Retrieved from http://ies.ed.gov/funding/ncer_rfas/readwrite.asp

Institute of Education Sciences (2010b). *What works clearinghouse*. USDOE/IES, Washington, DC. Retrieved from .http://ies.ed.gov/ncee/wwc/

Kochanek, J., & Clifford, M. (2011, April). *Refining a theory of knowledge diffusion among district administrators*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Kuhn, T. S. (1996). *The structure of scientific revolutions*. (3rd ed.). Chicago, IL: University of Chicago Press.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 30*, 4-16. doi:10.2307/1177052

Linn, R. L. (2005). Issues in the design of accountability systems. In J. L. Herman & E. H. Haertel (Eds.), *Yearbook of the National Society for the Study of Education: Vol. 104, Issue 2, Part 1. Uses and misuses of data for educational accountability and improvement* (pp. 78-98). Hoboken, NJ: Wiley-Blackwell.

Marchand-Martella, N. E., Slocum, T. A., & Martella, R. C. (2003). *Introduction to Direct Instruction*. Columbus, OH: Allyn & Bacon.

Meyer, L. A. (1984). Long term effects of the Direct Instruction Project Follow Through. *The Elementary School Journal, 84*, 395-407. doi:10.1086/461371

Murnane, R., & Willett, J. (2011). *Methods matter: Improving causal inference in educational and social and social science research*. New York, NY: Oxford University Press.

National Assessment of Educational Progress. (2009). Grade 12 national results. Retrieved from http://nationsreportcard.gov/reading_2009/gr12_national.asp

National Institute for Direct Instruction (2010a). *Research*. Retrieved from http://www.nifdi.org/15/research and http://www.nifdi.org/data.htm

National Institute for Direct Instruction (2010b). *NIFDI research reports*. Retrieved from http://www.nifdi.org/15/research and http://www.nifdi.org/data.htm

Popham, W. J. (2008). *Transformative assessment*. Alexandria, VA: ASCD.

Raudenbush, S. W., & Byrk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S.W., Bryk, A.S, & Congdon, R. (2004). *HLM 6 for Windows* [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Reeves, D. B. (2008). Looking deeper into the data. *Educational Leadership*, *66*(4), 89-90.

Ronka D, Lachat, M. A., Slaughter, R., & Meltzer, J. (2008). Answering the questions that count. *Educational Leadership*, *66*(4), 18-24.

Stebbins, L.B., St. Pierre, R.G., Proper, E.C., Anderson, R.B., & Cerva, T.R. (1977). *A planned variation model. Vol. IV-A Effects of Follow Through models.* U.S. Office of Education, Washington, DC.

Sidman, M. (1960). *Tactics of scientific research.* New York, NY: Basic Books.

Slavin, R. E. (2008a). Evidence-based reform in education: Which evidence counts? *Educational Researcher, 37*, 47-50. doi:10.3102/0013189X08315082

Slavin, R. E. (2008b). Perspectives on evidence-based reform in education. What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*, 5-14.

Sloane, F. (2008). Through the looking glass: Experiments, quasi-experiments, and the medical model. *Educational Researcher, 37*, 41-46.

Stebbins, L. B., St. Pierre, R. G., Proper, E. C., Anderson, R. B., & Cerva, T. R. (1977). *Education as experimentation*: A planned variation model (Vol. IV-A). Cambridge, MA: Abt Associates.

Stockard, J. (2008). *The What Works Clearinghouse beginning reading reports and rating of Reading Mastery: An evaluation and comment.* Technical Report 2008-04, National Institute for Direct Instruction, Eugene, OR.

Stockard, J. (2010). An analysis of the fidelity implementation policies of the What Works Clearinghouse. *Current Issues in Education*, 13(4). Retrieved from http://cie.asu.edu/ CIE

Stuart, E. A. (2007). Estimating causal effects using school-level data sets. *Educational Researcher*, 36, 187-198.

Tarver, S. (2004). Myths about Direct Instruction and research that refutes those myths. Retrieved from http://www.schoolinfosystem.org/pdf/112004/ditarver.html).

Vitale, M. R., Romance, N. R., & Kaniuka, T. S. (2011, April). *Methodological enhancements to school accountability for overcoming barriers to school reform: Implications for educational leaders.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

What Works Clearinghouse (2010). *Beginning reading.* Retrieved from http://ies.ed.gov/ncee/wwc/reports/Topicarea.aspx?tid=01

Author Note

Michael R. Vitale, Department of Curriculum and Instruction, East Carolina University, Theodore S. Kaniuka, Department of Educational Leadership, Fayetteville State University:

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305G04089 and by the National Science Foundation through Grant REC 0228353 to Florida Atlantic University. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education, or the National Science Foundation.

Correspondence concerning this article should be addressed to Michael R. Vitale, Department of Curriculum and Instruction, East Carolina University, Greenville, NC 27858. Email: vitalem@ecu.edu