The Journal of Special Education
http://sed.sagepub.com/

Designing High-Quality Research in Special Education : Group Experimental Design

Russell Gersten, Scott Baker and John Wills Lloyd

J Spec Educ 2000 34: 2
DOI: 10.1177/002246690003400101

The online version of this article can be found at:
http://sed.sagepub.com/content/34/1/2

Published by:
Hammill Institute on Disabilities

and
SAGE
http://www.sagepublications.com

Designing High-Quality Research in Special Education:
Group Experimental Design

Russell Gersten and **Scott Baker,** *University of Oregon, Eugene Research Institute*
John Wills Lloyd, *University of Virginia*

This article discusses critical issues related to conducting high-quality intervention research using experimental and quasi-experimental group designs. As researchers have learned more about teaching and learning, intervention studies have become more complex. The research community is struggling with ways to sensibly negotiate a balance between rigorous research designs that satisfy traditional laboratory standards of quality and interventions that are complex and flexible enough for conducting research in the real world of classrooms and schools. Rather than organizing the discussion around a list of resolute research standards, we weigh the pros and cons of making the many difficult choices involved in conducting intervention research. Our goal is to convey the sense that good designs must involve a series of balances and compromises that defy easily categorized solutions. Among the controversial areas discussed are the importance of defining the nature of the independent variable, the value of measuring implementation, and the improvement of the quality of quasi-experiments.

In an era when the public cries out for more information about research-based practices (Billups, 1997; Gersten & McInerney, 1997; Kornblet, 1997), it is indeed ironic that the number of intervention research studies investigating the effectiveness of special education instructional approaches is at one of its lowest levels in 30 years. Scruggs and Mastropieri (1994) have provided insights into why this shortage exists:

> When considering intervention research with students with learning disabilities, one is initially struck by the paucity of such research relative to other types of research in learning disabilities. Why does this relative scarcity of intervention research exist? Forness and Kavale (1987) argue that early special education research was conducted by psychologists who were more comfortable evaluating psychological characteristics of exceptional populations than evaluating the effectiveness of classroom interventions. As a result, "special education interventions did not evolve as completely as they should" (Forness & Kavale, 1987, p. 7). It is also possible that fewer intervention research studies are conducted because such studies are simply more difficult to design and execute and more costly in terms of necessary resources. (pp. 130–131)

In other words, much of the research on special education populations has focused heavily on *describing* psychological attributes and levels of educational achievement and has tended to avoid research on the effects of interventions. Scruggs and Mastropieri (1994) suggested that this evolution is due, in part, to the difficulties inherent in designing school-based intervention studies. This is particularly true when the goal is to use group design methods for evaluating the impact of instructional interventions.

Studies using group designs remain the primary means for assessing whether educational interventions have beneficial effects on students. Although qualitative studies can provide valuable insights into the process of change and enhance understanding of facets of teaching and learning, experimental group designs remain the most powerful method available for assessing intervention effectiveness (Cook & Campbell, 1979; Gall, Borg, & Gall, 1996; Krathwohl, 1993; Slavin, 1999; Vockell & Asher, 1995).

As concern about the effectiveness of special education increases, so too does the field's need for valid and reliable evidence about best practice. For the past 5 years, a working group composed of more than 20 special education researchers assembled by the Office of Special Education Programs (OSEP) has met to address this issue. The group's discussions have produced stimulating dialogue, occasional

Address: Russell Gersten, Eugene Research Institute, 132 E. Broadway, Ste. 747, Eugene, OR 97401

heated debate, and successively clearer conceptualizations of major issues and roadblocks to the conduct of compelling field research involving students with disabilities. Some participants have argued that the trend over the past 20 years has been toward increasing the external validity of findings by conducting more research in real classrooms or community settings. Unfortunately, this trend has served to compromise the technical standards of group research, which is troublesome at a time when producing credible, valid evidence regarding the impact of educational interventions is particularly important.

What can be done to generate more compelling evidence about effective educational practices for students with disabilities? Maintaining a focus on conducting intervention research in real school settings is imperative, and we believe that one of the best ways to support this objective is to develop a consensus on how to design more rigorous and high-quality applied research studies. That is a major goal of this article.

The initial purpose of the OSEP meetings was to develop standards for improving the quality of group intervention research. Editors of journals and authors of contemporary essays about intervention research (e.g., Carnine, 1995; Gersten et al., 1998; Graham & Harris, 1994) reinforced our concern about the dwindling number of intervention studies.

The focus was placed squarely on *group* intervention research, not only because of its methodological preeminence for comparing different instructional approaches and techniques, but also because there was widespread concern among our group's members that the number and quality of group intervention designs needed to be increased and improved.

The design of a good study always incorporates balances and compromises. One reality of conducting experimental research in special education is that, in general, the shorter the length of the study, the more precision there is in attributing changes in student performance (i.e., the dependent variable) to the instructional intervention (i.e., the independent variable). However, a critical empirical question must always be whether the effects of the intervention will persist over time—that is, do the effects last beyond a very brief period of time, or beyond the duration of the study?

As we have learned more about the effectiveness and nuances of specific instructional approaches and strived to determine their effectiveness in real classroom settings, intervention studies have naturally become more complex and longer. Consequently, it has become more challenging to conduct tightly controlled experimental research. More research also is being conducted by teacher–researchers in an effort to increase the ecological validity of findings (e.g., Chamot, Keatley, & Mazur, 1999; Englert, Raphael, Anderson, Anthony, & Stevens, 1991; Englert & Tarrant, 1995; Wong, Butler, Ficzere, & Kuperis, 1997). These participating teachers are frequently given discretion in how to implement the principles of teaching and learning that underlie the study. This type of collaborative research challenges traditional concepts of replication as well as the conventions for reporting procedures that evolved from experimental psychology.

We want to encourage applied research in school and community learning contexts, using research designs that address the realities of school practice. Some traditional standards promulgated in textbooks must be adjusted if meaningful intervention research is to be conducted in these settings. Judging how and when to modify standards developed from laboratory psychology studies without compromising the integrity of designs is critical. Making such issues explicit and illuminating principles that help create more valid research in school and community settings is a major goal of this article.

In more formal (if somewhat antiquated) terms, the unavoidable trade-off between internal and external validity must be actively addressed. It is equally important, however, to understand common, predictable pitfalls in design and execution (e.g., failing to adequately assess the intervention implemented against the intervention desired—a discrepancy that can be so severe that the assertions made by the researchers are not credible; Cooper & Hedges, 1994; Gersten & Baker, in press; National Center to Improve the Tools of Educators, 1998; Swanson & Hoskyn, 1998).

The reality is that an increasing number of quasi-experimental designs are being employed that use intact classrooms and sometimes intact schools (Hunt & Goetz, 1997; Slavin & Madden, 1995) as the means of assigning participants to conditions. These studies are developed in response to real-world concerns about the education of students with disabilities. Thus, in this article, we address issues related to both experimental and quasi-experimental designs.

Essentially, the purpose of this article is to discuss critical issues related to conducting high-quality intervention research using experimental and quasi-experimental designs that compare outcomes for different groups of students. We hope to inform new researchers of these issues and share the craft knowledge provided by the participants in the OSEP group. We articulate how we, as a research community, sensibly negotiate a balance between design components that satisfy laboratory standards and those that reflect the complexities of real-life classroom teaching. Above all, intervention research must provide reliable and clear answers to meaningful questions.

Overview

We want to encourage more research on the effects of interventions for groups of students using well-designed methods. Group designs have been attacked both by advocates of single-subject research (in the 1970s and early 1980s) and by advocates of qualitative research (in the past decade). The wave of critiques of group experimental designs for studying research on teaching, no matter how well intentioned and thoughtful (e.g., Ball, 1995; Kennedy, 1997; Richardson, 1994), has taken a toll on what remains our most powerful tool for under-

standing the effectiveness and impact of instructional interventions and for influencing educational policy for students with disabilities.

Designing high-quality experimental comparisons requires sophisticated attention to research methodologies. Rather than presenting an exhaustive treatise on every facet of research design, we offer recommendations about several key aspects and discuss common and thorny issues that persistently recur. Our goal is not to reiterate the principles of experimental research design as found in standard texts. Instead, we hope to ground key principles in the realities of current classroom practice, using the expertise and experiences of the active instructional researchers who engaged in the discussion groups over a 5-year period.

Of course, the members of the OSEP group acknowledged the virtual impossibility of meeting all the criteria for group research discussed in standard textbooks. However, we believe that it is possible to design studies in applied settings that are sufficiently flexible while maintaining the rigor to provide valid information on the extent to which theories about teaching and learning and social and cognitive growth lead to instruction that enhances student learning.

We address several controversial areas within group design, including topics on which group members provided interesting insights and perspectives but failed to reach consensus. It seemed important to present the diverse views on these dilemmas, provide rationales, and explain some of the alternative solutions that members of the group proposed. Among the controversial issues discussed are defining the nature of the independent variable, assessing intervention implementation, selectively using an alternative to formal experimental research, assigning students to treatment conditions, using quasi-experiments, selecting dependent measures, and conducting replication studies. These concepts serve as major organizers for this article.

Operationalizing the Independent Variable

Precise descriptions of independent variables are crucial to furthering our knowledge base. Most texts treat the operationalization of instructional methods as a fairly routine activity, as a way of merely checking for fidelity (i.e., checking whether teachers are implementing the approach in the fashion that the researchers specified). However, the task is far more intricate than one might think (Kennedy, 1997; Kline, Deshler, & Schumaker, 1992). Researchers often have a very good conceptual sense of what they would like to see during instruction, but describe only half-formed images of the types of specific actions and behaviors that constitute implementation on a day-by-day, minute-to-minute basis (Kennedy, 1991).

The difficulty of operationalizing an instructional approach is exacerbated by efforts to design interventions that work in the real world. For example, teachers are increasingly included in creating and operationalizing the instructional approaches being investigated in studies (Chamot et al., 1999; D. Fuchs & Fuchs, 1998; Klingner, Vaughn, & Schumm, 1998). This practice necessarily tends to result in more flexible specifications of instructional components than if university researchers were to develop the details of instruction on their own. Also, to increase the prospect that instructional interventions will be applicable to the realities of classroom practice, high degrees of teacher autonomy in making decisions about how to deliver interventions are becoming increasingly common (Gersten, Vaughn, Deshler, & Schiller, 1997; Malouf & Schiller, 1995; Scanlon, Schumaker, & Deshler, 1994).

It is important to recognize that instructional labels may vary significantly from study to study and that, within a study, the *intended* intervention may only marginally resemble what is actually implemented. In other words, the names assigned to instructional interventions can be quite misleading (Swanson & Hoskyn, 1998). Across studies, for example, classwide peer tutoring may be implemented to address aspects of reading fluency in one study and comprehension in another. Similarly, in some versions of peer tutoring, the students are provided with specific guidelines for helping out their partners when they experience difficulties; in others, students are left to their own devices for providing assistance. This variation of components occurs in virtually all areas of instructional research, be it cognitive strategy instruction in writing (Graham & Harris, 1989), direct instruction research in mathematics (Carnine, Jones, & Dixon, 1994; Swanson & Hoskyn, 1998), or situated cognition (Bottge & Hasselbring, 1993).

Only by carefully analyzing a study and what transpired during its lessons is it possible to understand which elements led to specific outcomes. Of course, in reality, each study can only imperfectly describe or allude to the myriad of details that constitute the precise nature of the independent variable.

Yet, through a combination of programmatic research, independent replications, and component analysis, which is at the heart of high-quality research syntheses and meta-analyses (e.g., Rosenshine & Meister, 1994; Swanson & Hoskyn, 1998), we can begin to discern patterns of practices that lead to improved outcomes for students.

Precision in operationalizing the independent variable becomes increasingly important as replication studies become more refined and synthesis procedures such as meta-analysis become more common. In fact, the purpose of research syntheses is to "discover the consistencies and account for the variability in similar-appearing studies" (Cooper & Hedges, 1994, p. 4). When analyzing similarities and differences among independent variables (i.e., instructional interventions) across multiple studies, or when trying to determine the effect that subtle changes in instruction might have on learning outcomes, precise descriptions of instruction are critical.

Swanson and Hoskyn (1998) in their meta-analysis of the impact of various instructional approaches on students with learning disabilities illustrated the importance of precise descriptions of instruction. They concluded that two

approaches—*direct instruction* and *strategy instruction*—were almost equally effective in enhancing learning. Both approaches had consistent, moderately strong effects, with virtually identical effect sizes (.68 for direct instruction and .72 for strategy instruction). However, Swanson and Hoskyn's most meaningful insight was their observation that there was significant overlap in the way the two constructs were operationalized. They concluded that classifying types of teaching with terms such as direct instruction and strategy instruction was problematic—more fine-grained terminology and descriptions needed to be used.

Regardless of the approach, be it direct instruction (Losardo & Bricker, 1994), anchored instruction (Kline et al., 1992), or strategy instruction (Sawyer, Graham, & Harris, 1992), the details of the independent variable directly influence which research questions can be answered. Thus, it is critical to know how the instructional intervention is actually delivered. We will have more to say about implementation issues later.

The Gap Between Conceptualization and Execution

Slippage between the conceptualization of a study and its execution is one of the most common problems in applied research, and unfortunately this problem is sometimes so severe that the outcomes are uninterpretable. Of course, an important part of any good intervention study is the serious consideration of rival explanations in accounting for outcomes. Conducting applied research does not absolve researchers of their responsibility to control for confounding variables (e.g., minutes of instruction in treatment and comparison groups, or overall teaching effectiveness of teachers assigned to experimental and comparison groups) and to investigate the way variables are operationalized during the study (e.g., degree of support provided students to identify and explicate a theme in a story; J. P. Williams, Brown, Silverstein, & deCani, 1994).

Researchers should provide detailed descriptions of interventions, providing enough information for their replication. Any in-depth examination of implementation by the use of audiotapes or sophisticated observational systems such as the Code for Instructional Structure and Student Academic Response (CISSAR; Greenwood & Delquadri, 1988) or The Instructional Environment Scale (TIES; Ysseldyke & Christenson, 1987) also greatly enhances the quality of a study. Providing actual transcripts of lesson segments and instructional interactions also provides rich insights into the true nature of the intervention (e.g., Echevarria, 1995; D. Fuchs, Fuchs, Mathes, & Simmons, 1997).

A group intervention study by Lovett et al. (1994), comparing direct instruction and strategy instruction with middle school students with reading disabilities, is a good example of providing a precise specification of the independent variable. Both instructional methods included intensive word identification instruction and procedures for identifying unknown words. The two conditions were clearly differentiated on a number of salient dimensions, including conceptual underpinnings, teaching strategies employed, instructional materials, and what students were required to do. This was a complex study, yet it also contained clear classroom applicability. It has furthered our knowledge base on effective reading instruction for students with learning disabilities precisely because theories and their instructional components were defined in a fashion that could be used for replication or easily coded in a research synthesis such as meta-analysis.

Measuring the Independent Variable

Assessing the implementation of educational interventions and approaches has a fascinating history and is one of the most interesting and complex issues in applied educational research. In the 1970s and 1980s, an era of many large-scale evaluations, the importance of assessing the extent to which an educational intervention or approach was actually implemented was stressed in texts and articles. Charters and Jones (1974), for example, chastised the field for evaluation of "non-events" (i.e., evaluation of programs that, for one reason or another, were not actually implemented). This criticism led to a rash of studies of how various innovative instructional approaches were actually implemented in classrooms (e.g., Good & Grouws, 1977; Leinhardt, 1977; Stallings, 1975). These concerns also spawned the development of sophisticated systems for assessing and understanding implementation, such as the Concerns-Based Adoption Model developed by Hall and Loucks (1977).

To indicate the importance of implementation in large-scale intervention research, we refer to a study by Hasselbring et al. (1988) on assessing the impact of a laser disc instructional program in mathematics. These researchers found that, although the program was intended for daily use, many teachers used it only once each week. In other words, the researchers were essentially evaluating a nonevent. In contrast, a study of laser disc use by Woodward and Gersten (1992), in which careful monitoring of implementation occurred, revealed daily use of the program by all teachers. Clearly, these results from two studies of a similar intervention must be interpreted cautiously in light of marked differences in implementation. Interpreting the differences in the two studies in the context of their independent variables would reveal that the studies were nominally similar but functionally very different.

There are many stumbling blocks to valid assessment of implementation. Several factors have curtailed the major advances in implementation measurement that were made in the 1970s and 1980s. The first was a drastic reduction in the number of large-scale evaluations. Second, implementation research—especially implementation research that requires direct observation of classroom interactions—is expensive. In fact, our experience suggests that the assessment of implementation can be as costly as the administration of pre- and postintervention measures. In the next sections, we provide a brief overview of the issues and current thinking on the topic of measuring the fidelity of implementation and assessing the

extent to which a comparison group may also be implementing critical components of the intervention.

Implementation Fidelity

It is essential that researchers gather and report core implementation fidelity information, such as the level of training provided to participants, the length of lessons, whether critical aspects of teaching were in fact implemented in each room, the amount of time each day dedicated to the intervention, and so forth (see e.g., Kelly, Gersten, & Carnine, 1990). Unfortunately, reporting even minimal assessments of implementation is rare (Gresham, Gansle, Noell, Cohen, & Rosenblum 1993; Peterson, Homer, & Wonderlich, 1982; Troia, 1999). For example, a review by Gresham et al. (1993) of 181 experimental studies published between 1980 and 1990 revealed that only 14.4% systematically measured and reported treatment integrity data. In reviewing experimental studies involving phonemic awareness interventions, Troia found that only 5 of 39 studies (13%) reported treatment fidelity data. It was interesting that 3 of these 5 studies were by the special education research team of O'Connor, Jenkins, Slocum, and their colleagues (i.e., O'Connor, Jenkins, Leicester, & Slocum, 1993; O'Connor, Jenkins, & Slocum, 1995; Slocum, O'Connor, & Jenkins, 1993).

In other words, the data suggest that even minimal measures of implementation are not a common part of researchers' methodological procedures for group intervention research. Consequently, it can be argued that any type of consistent implementation assessment would be an improvement over the current state of affairs. However, it is important to note the value of different types of implementation assessment procedures. If assessing implementation is an afterthought done with only marginal rigor, the knowledge base on implementation will not be advanced to any notable degree. With minimal fidelity checklists, an impartial observer may record the occurrence of the most central aspects of the intervention and determine if the experimental and comparison groups receive different instruction. Do students work in small groups and give each other feedback on their written essays? Does the teacher use the new history textbook during teacher-led history instruction?

More sophisticated rating forms are capable of advancing the knowledge base on implementation. In other words, they can help us understand the degree to which more subtle aspects of an intervention are implemented and the types of modifications that occur. They can be used to record the occurrence of certain types of instructional techniques, to note whether students are provided with models of proficient reading during their work with peers, and to record if students have the opportunity to verbally present the story map they have been working on.

For example, a rating form used for a type of classwide peer tutoring called peer assisted learning strategies (D. Fuchs et al., 1997) required an evaluator to observe a lesson and note not only whether the teacher followed all prescribed steps in the process but also whether students read in pairs, were awarded points, and used some kind of error correction strategy. Implementation rating forms for SRA Reading Mastery (Gersten, Carnine, Zoref, & Cronin, 1986) also examined a wide array of critical teacher and student performance variables.

A limitation of implementation fidelity checklists is that the more elusive aspects of superior implementation (e.g., quality of examples used, type of feedback provided), which often are at the heart of complex interventions, may not be captured by a checklist or rating form. A deeper level of understanding implementation issues is more likely to occur when more qualitative observation procedures are used, combined with interviews with participating teachers to gain their perspectives.

One reason for the decline in studies of implementation was a convergence of findings (e.g., Cooley & Leinhardt, 1980) indicating that generic features of effective instruction transcended any given intervention. Initially, this was an unanticipated finding, based on studies that attempted to link the degree of implementation with student outcomes. As researchers began to probe in increasing depth the precise features of direct instruction (Gersten, Carnine, & Williams, 1982), individualized instruction (Leinhardt, 1977), and other interventions geared toward low-achieving students (Stallings, 1975), they found that common features among these models—such as the amount of academic engaged time, the continuous monitoring of student progress, and the quality of feedback provided to students when they encountered difficulties—were as important as the nominal label of the intervention.

Graham and Harris (1994) argued for a renewal of research involving studies of implementation in special education. They urged researchers to "assess . . . the processes of change as related to both intentions and outcomes . . . determining the relative contributions of instructional components and the variables responsible for change" (p. 151).

In the past decade, qualitative studies of implementation, often using discourse analysis, have begun to appear in the literature (e.g., Ball, 1990; L. S. Fuchs, Fuchs, Bentz, Phillips, & Hamlett, 1994; Gersten, 1996; Palincsar & Brown, 1984; S. R. Williams & Baxter, 1996). These studies use analyses of audiotapes or videotapes to capture the nuances of implementation. This use of selected verbatim transcripts has substantially increased our understanding of what really happens when a class uses classwide peer tutoring, content-area sheltered instruction, reciprocal teaching, or instruction based on guidelines from the National Council of Teachers of Mathematics.

Another essential aspect of implementation fidelity is the time frame for conducting implementation checks. When reporting implementation, it is helpful to specify the periods

when implementation checks were conducted. At minimum, implementation checks should occur at the beginning of a study, a few weeks later to verify corrections, and again midway to late in the study. Assessing implementation fidelity allows the research team to check for slippage (i.e., unplanned deviations from the intended instructional approach) later in implementation.

Contemporary investigations of implementation have also begun to assess teachers' understandings of the underlying thinking behind an intervention. These investigations have allowed for a deeper level of understanding of how teachers adapt interventions and of the extent to which these adaptations have integrity. In other words, the specific components of an intervention may be modified by teachers to better fit into their classrooms. These modifications may result in learning outcomes that are equally effective, less effective, or more effective than the original intervention. Our methods of assessing implementation should attempt to evaluate the effectiveness of these teachers' modifications and changes. Interviews with teachers—especially ones that focus on rationales for specific options—can greatly enhance our understanding of the feasibility of the intervention. They also may clarify what we mean by implementation with integrity (Graham & Harris, 1994).

In summary, assessment of implementation is complex and frequently neglected by researchers. Assessing an implementation is often difficult and expensive to do properly. Also, the special education field has devoted insufficient attention to the topic in evaluating the quality of research designs.

The Nature of the Comparison Group

The following comment about comparison groups illustrates one clear way in which the quality of intervention research can be improved:

> Interventions are best evaluated relative to credible comparison conditions. . . . One way to improve the design of credible alternative conditions is communication (and potentially even collaboration) with scientists who are well informed about the alternative interventions. To the extent that intervention researchers perceive studies to be horse races—that are either won or lost relative to other interventions—constructive communication and collaboration with workers representing alternative interventions is unlikely. (Pressley & Harris, 1994, p. 197)

One of the least glamorous and most neglected aspects of research is describing and assessing the nature of instruction in the comparison group. Yet, to understand what an obtained effect means, one must understand what happened in the comparison classrooms. This is why members of the research team also should assess implementation in comparison classrooms. At a minimum, researchers should examine comparison classrooms to determine what instructional events are occurring and what professional development and support is provided to teachers. Factors to assess include possible access to the curriculum/content associated with the experimental group's intervention, time allocated for instruction, and type of grouping used during instruction (Elbaum, Vaughn, Hughes, & Moody, 1999).

Based on prior findings and on the knowledge base, some other questions that should be explored include (a) Did the two conditions differ markedly in the amount of feedback available to the learners? (b) How many demonstrations and practice opportunities were provided to each group? and (c) Were the learners in both groups equally likely to receive personal encouragement for persisting in solving problems?

Perhaps the most serious threats to interpretation are related to what Swanson and Hoskyn (1998) labeled one of the major problems in special education research—confounding teachers with an intervention approach (i.e., classroom by treatment confounds). Scruggs and Mastropieri (1994) also noted the seriousness of this problem: "Assessment of relative treatment efficacy is extremely difficult from a design that essentially confounds treatments with classrooms" (p. 133). In addition to issues of overall teaching quality and differences between experimental and comparison group teachers, Scruggs and Mastropieri suggested that teacher enthusiasm is also a confound. They explained that teacher enthusiasm in the experimental group has been associated with strong impacts on academic and social learning, even when the same curriculum and instruction were employed in the comparison group. If one teacher instructs the experimental group and another teacher instructs the comparison group, differences in student performance could be attributed in part to different levels of teacher enthusiasm rather than to the independent variable being studied.

We must ensure that the quality of teaching is similar across comparison conditions. One means of accomplishing this goal is to *counterbalance* teachers across both treatment and comparison conditions (Dimino, Gersten, Carnine, & Blake, 1990; Echevarria, 1995). Moreover, it is critical that the research team avoids relying on a single teacher to deliver the experimental approach. Researchers should develop procedures to ensure that teaching quality is basically equivalent across conditions. Swanson and Hoskyn (1998) found that using a single teacher to implement an instructional approach invariably inflated effect sizes. They concluded that these studies may well confound teacher quality with instructional approach. It is also valuable to assess teacher effects on a post hoc basis, using typical analysis of variance procedures (Dimino et al., 1990). Figure 1 provides a summary of critical issues to consider in trying to better define and operationalize the instructional approach.

1. Avoid the "nominal fallacy" by carefully labeling and describing independent variables.
2. Search for unanticipated effects that may be produced by the instructional intervention.
3. Address assessment of implementation using standard checklist procedures as well as in-depth methods analysis.
4. Carefully document and assess what happens in comparison classrooms or other settings.

FIGURE 1. Major recommendations for defining and operationalizing the instructional approach.

Designs for In-Depth Understanding of Teaching and Learning

Probing the Nature of Instruction

In the past 5 years, an increasing number of instructional researchers have suggested that alternatives to traditional experimental or quasi-experimental designs be used in research on teaching and learning. The central problem with experimental designs is that efforts to control, manipulate, and understand a narrow and precisely defined independent variable rarely result in a deep understanding of the realities of classroom implementation. These designs also do not reveal how the independent variable, interacting with other aspects of the instruction, contributes to the learning of complex content. One partial solution is to conduct flexible studies that allow for a deeper understanding of what an independent variable might actually look like, given the realities of classrooms, in advance of conducting formal experimental and quasi-experimental studies. This can be the cornerstone of programmatic research.

Design Experiments/Formative Experiments

Calls for the increased use of such alternative designs have originated from a range of disciplines (e.g., technology, science education, cognitive science, reading comprehension), and the nature of these proposed alternatives is remarkably similar. These studies are increasingly visible in the literature and go by a variety of names: design experiments (Brown, 1992), formative experiments (Newman, 1990; Reinking & Pickle, 1993), or developmental studies (Gersten, 1996; Gersten & Baker, 1997). We do not envision such design experiments as supplanting experimental research. Rather, design experiments may be a useful tool for conducting research on newer, less well-defined topics such as technology applications, integration of technology with instruction, and studies of teacher change. Newman (1990) provided a helpful conceptualization of what a design experiment is and how it unfolds:

> [The] plan for implementation . . . conceived at the beginning is seen as [a] first draft, subject to modification during the experiment. Through systematic investigation, the researcher(s) observe and document factors that inhibit or enhance implementation of the intervention and achievement of the pedagogical goal. (p. 264)

In other words, Newman (1990) suggested that, in design experiments, researchers begin with specifying the desired goal. Then, based on student performance data and responses from teachers who are implementing the intervention, the intervention should be continually adjusted to reach the goal. This modification process is necessary because, with less-tested instructional methods, it is not possible to select the appropriate length of intervention or the most valid, relevant dependent measures until one is actually immersed in a classroom.

It is important to note that advocates of design experiments (Beck, McKeown, Sandora, Kucan, & Worthy, 1996; Brown, 1992; Reinking & Pickle, 1993; Richardson & Anders, 1998) have not argued for total abandonment of traditional quantitative measures of student learning. Rather, they have argued for a mix of qualitative and quantitative methods in the context of a design experiment.

Although these designs are still in the early stages of development, we believe they hold promise for those involved in instructional research. The limitations of conventional designs for conducting research on classroom teaching and learning raise significant issues for the entire educational research community.

This point was made in a self-deprecating fashion by Reinking and Pickle (1993). Their description captured the frustrating experience of many who conduct applied instructional research in classrooms. In their study of computer use and literacy, they described the unanticipated changes and compromises they needed to make for successful implementation:

> As the school year progressed, we found ourselves in a seemingly endless cycle of compromises that threatened the control required in a true experiment . . . . Each compromise seemed like a defeat in a war we were quickly losing . . . . Our need to maintain control of extraneous variation was a barrier to finding and understanding the most relevant aspects of implementing the intervention and the effects it might have on the educational environment. (pp. 266–267)

The problem of maintaining control over extraneous variables has been exacerbated as the length of interventions

has extended from days to months and as there has been a shift away from easy-to-measure skills, such as math computation, to more complex problem-solving skills characteristic of real-world classrooms.

As researchers have attempted to study changes in teaching, the drawbacks of traditional designs have become apparent. Richardson and Anders (1998) explained that studies of implementation "almost invariably take unexpected twists and turns" (p. 25) as researchers examine teachers' adaptations of instructional practices developed by researchers. They noted how formal quasi-experimental designs, or studies where the variables have been too precisely defined in advance, can inhibit understanding of the process being investigated. Based on our own experience and reading of research on the change process, we believe this is a valid point.

Richardson and Anders (1998) urged that, while studying a complex issue such as teacher change, researchers should concurrently collect both objective data and more subjective data and be "open to surprises and new understandings in learning about the process and its results" (p. 94). Ideally, research questions, emerging research issues, and results should be investigated "in a way that allows the reader to continue thinking about the process, data, and consequences" (p. 94).

Brown's (1992) extensive use of instructional interviews to assess students' understandings of scientific content demonstrated the advantages of flexibility in investigating aspects of student learning that occur in new areas of inquiry. She began by asking a child a series of basic questions dealing with factual or declarative knowledge of key scientific concepts, such as the food chain or photosynthesis. If the student was unable to answer, Brown provided examples or prompts. If the student appeared to know the concepts, depth of understanding was probed with a series of examples and counterexamples.

Brown (1992) argued that only this in-depth probing through a range of examples enabled her research team to understand what students really learn about scientific concepts, and which concepts require additional discussion. In Brown's words, these dynamic assessments "allow us to track not only retention of knowledge, but also how fragile [or] robust it is and how flexibly it can be applied" (p. 159). This information is used to refine aspects of the independent variable by conducting a formal experiment and to sharpen the sensitivity of dependent variables to assess what is really being learned.

In a similar way, Pressley and El-Dinary (1997) described how their open-ended design experiments of comprehension strategy instruction permitted them "to construct a far more complete model of comprehension strategies instruction. . . .The insights . . . gained from [the design experiments] permitted the design of a quantitative, comparative study of teacher-implemented comprehension strategies instruction that, we believe, was more realistic than previous studies of the effects of comprehension strategies instruction" (Brown, Pressley, Van Meter, & Schuder, 1996, as cited in Pressley & El-Dinary, 1997, p. 488). Interventions based on design experiments tend to be more dynamic and responsive to the complexities of classroom environments than those developed in isolation at a university or research institute.

We believe that design experiments can and should be a critical tool in refining innovative instructional practices in real classroom environments and formally documenting their effects. Design experiments are particularly useful in gaining an in-depth understanding of the relationship between specific aspects of instruction and learning on a variety of performance measures and the nature of effective adaptations for students with disabilities.

Selecting, Describing, and Assigning Students to Conditions

Identifying and describing the group being studied is central to designing and implementing quality group design. Although research textbooks often describe this process as relatively straightforward, it is actually quite complex for many reasons that are explored in this section.

Researchers conducting studies with special populations are faced with the extraordinary challenge of identifying populations that are sufficiently homogeneous to constitute a group and yet large enough to provide adequate power for group comparisons. Too often, intervention studies with special populations yield nonsignificant results because there are too few participants in each group. As a general rule, more is better. When there is an adequate body of prior research on the measures and sample to be used in a study, power analyses should always be conducted to help make decisions concerning the minimal sample size necessary and the number of comparison conditions that are truly viable. When researchers approach new areas where no such data exist, the old aphorism "20 is plenty" (i.e., 20 students per condition is adequate) remains reasonable advice. Rarely will sample sizes of 12 to 15 be adequate unless the anticipated effects are extraordinarily strong.

Studies that contrast slight variations on instructional interventions are increasingly being conducted (e.g., Bottge & Hasselbring, 1993; L. S. Fuchs, Fuchs, Kazdan, & Allen, 1999; Graham, MacArthur, & Schwartz, 1995; O'Connor & Jenkins, 1995) as opposed to studies in which experimental treatments are compared to no-treatment controls. Too frequently, researchers forget that minimally different treatments require larger sample sizes to uncover effective interventions.

An alternative to increasing power is to increase the homogeneity of the groups involved in the study. There are obvious trade-offs with this technique, the foremost limitation being a significant decline in the generalizability of findings.

More Thorough Sample Descriptions

The importance of adequately describing samples increases with the growing emphasis on synthesizing research findings. Swanson and Hoskyn (1998) noted, for example, how read-

ing studies involving students with learning disabilities showed larger effects with a sample of students at or below the 16th percentile. McKinney, Osborne, and Schulte (1993) found that attention deficits had a major impact on how well students responded to instructional intervention.

The Research Committee of the Council for Learning Disabilities (Rosenberg et al., 1994) noted that available descriptions of individuals with disabilities in research reports are vague and inconsistent. Inadequate descriptions of participants make it difficult at best, and sometimes impossible, to evaluate research findings or to replicate studies. These problems are particularly acute for studies involving students with learning disabilities (LD) because of the complexity and variability of definitions used to determine their eligibility for special education. The committee recommended that the following variables be used for describing students with disabilities: (a) gender, (b) age, (c) race or ethnicity, (d) level of English language development, (e) socioeconomic status, (f) achievement levels on standardized tests, and (g) intellectual status of the participants. In research with small sample sizes (fewer than 10 participants), a more thorough description is warranted (see Klingner & Vaughn, 1996, as a model).

Correlations between pretest and posttest measures should be routinely calculated for experimental and comparison groups. These correlations help researchers begin to understand which student characteristics are most likely to predict success or failure with a given approach. For populations as diverse as those typically involved in special education research, these secondary analyses can be extremely important. This is especially true for disability categories, such as learning disabilities, where comorbidity is common. McKinney et al. (1993) found, for example, that academic outcomes were quite different for students with LD depending on the presence or absence of attention problems, even if students began the study with equivalent pretest scores.

Information on subtypes of disabilities and on differential impacts of instructional interventions can be very valuable. However, currently there is a major controversy in the field of reading disabilities as to the merits of focusing on school-identified LD samples. Lyon and Moats (1997), among others, argued that research with school-identified samples of students with LD is not valid and that researchers should always address the full range of student abilities and disabilities or operationally define the reading disabilities sample without reference to labels given by schools. On the other hand, researchers such as Fuchs and Fuchs have conducted numerous studies using school-identified samples of students with LD and demonstrated the impacts of various interventions on the students with learning disabilities, other low-achieving students, and average students (e.g., D. Fuchs et al., 1997). An advantage of using school-identified samples is that findings generalize to students with LD who are found in schools, making findings more useful and directly relevant for improving current practices. A recent meta-analysis documented that, across researchers and research studies, school-identified samples of students with LD invariably score significantly lower than samples without students with LD on all measures of achievement (D. Fuchs, Fuchs, Mathes, & Lipsey, in press).

Some researchers have used both types of procedures for sample selection depending on the question being asked. We conclude that both methods of sample selection—those employed by Fuchs and colleagues and those employed by researchers such as Foorman, Francis, Fletcher, Schatschneider, and Mehta (1998)—are valuable means for systematically improving the knowledge base on effective teaching strategies for students with learning difficulties.

Random Assignment Versus Random Selection: A Source of Confusion

Randomly assigning students to experimental and comparison groups is perhaps the signature characteristic of true experiments in intervention research. Occasionally, random *selection* is confused with random *assignment,* with problematic results. For survey research or demographic studies, random selection is critical. By using random selection, a research team can generalize their results to the population from which they drew their sample, thus addressing a particular study's *external* validity.

Although random assignment of students to treatment and comparison conditions is critical in intervention research, random selection is not. Randomly assigning students to experimental and comparison conditions provides greater certainty that differences between groups on outcome measures are the result of the treatment. This is primarily a matter of internal validity.

One way to achieve comparable groups—and high precision in a study—is to match pairs of students on a variable salient to outcomes in the study (e.g., reading fluency in a reading comprehension study) and then randomly assign one member of each pair to each condition (Cook & Campbell, 1979). Note that this procedure is quite different from finding a match for an experimental student after students in the experimental condition have been determined. Matching students on an important variable and then randomly assigning them to treatment and comparison conditions leads to a well-controlled true experiment. Finding a match for a student in the experimental group from a neighboring classroom or school leads to a quasi-experiment, with numerous potential problems outlined in detail by Campbell and Stanley (1963). (Note that data analysis must take into account the fact that these are dependent samples.)

Another viable approach is to stratify students on a salient variable, such as reading or writing ability as measured on standardized or performance measures, and randomly assign within each stratum (e.g., students with low scores, average scores, high scores). Both methods (as well as simple random assignments) are legitimate random assignment approaches, each with its own strengths and weaknesses.

When random assignment is not possible, quasi-experimental designs may be the only suitable alternative. However, the validity of inferences drawn from quasi-experiments will always be subject to question (Cook & Campbell, 1979).

Controversies Surrounding Quasi-Experimental Designs

Since the publication of Campbell and Stanley's (1963) classic monograph urging researchers to move into real-world settings, even if it means abandoning the random assignment of participants to treatment conditions, researchers have frequently used the quasi-experimental designs described by Campbell and Stanley. In true experiments, participants are randomly assigned to one of the intervention or treatment conditions. In quasi-experiments, researchers often use students from intact classes or schools as the intervention or treatment sample and try to find a relatively comparable group of students from other classes or schools to serve as the comparison sample.

In the most frequently used type of quasi-experiment, to explain or account for potential differences between the treatment and comparison groups, researchers typically assess students on a battery of pretest measures to ensure equivalence. If differences exist, analysis of covariance can be used to adjust statistically for these initial differences. Increasingly, in studies involving three or more data points, techniques such as growth curve analysis are used for the interpretation of individual pretest differences between students (Dunst & Trivette, 1994).

However, the use of quasi-experimental designs remains controversial, as a recent article by Greeno and the Middle School Mathematics Through Applications Projects Group (1998) noted, and as Campbell himself realized (Campbell & Erlebacher, 1970). Problems with quasi-experimental designs are particularly severe when pretest differences between treatment and comparison groups exceed one-half standard deviation (0.5 *SD*) on relevant criterion measures. In these cases, it is likely that students in the two groups come from different populations, which even under the best of circumstances is extremely problematic. It is also true that one can never use covariance on every possible variable on which the experimental and comparison groups may differ. It is always possible that an unknown variable differentiating the groups—and not the intervention—is actually responsible for the posttest results. For this reason, quasi-experiments can never really supplant true experiments. Our concern is that the increased use of quasi-experiments without adequate attention to important research design considerations has resulted in many studies so weak or compromised (e.g., due to clear bias in selecting students for the intervention group or to large initial differences at pretest) that it is unclear whether the data supported the researcher's assertions.

Researchers who conduct meta-analyses frequently exclude quasi-experiments unless their data reveal that no more than 0.25 *SD* separated experimental and comparison groups on salient pretest variables (National Center to Improve the Tools of Educators, 1998) and their researchers provided evidence that quality of teaching was not a confound. Even if a quasi-experiment meets these criteria, a quasi-experiment will never be an ideal substitute for a true experiment, regardless of how well it is designed and conducted and certainly no matter what the results are.

Because of the limitations of quasi-experiments, some researchers feel we need to encourage the increased use of true experiments—where participants are randomly assigned to conditions—and rely much less on quasi-experiments. Cases of intervention studies with random assignment of students to treatment groups and high ecological validity are present in the special education literature.

However, some members of the OSEP work group felt that, given the constraints of working in schools and clinics, true experiments were impossible to conduct on a large scale. Still others argued that researchers often failed to put sufficient energy into negotiating for random assignment. The work group concluded that, because quasi-experiments are a way of life for many researchers, standards for conducting quasi-experiments should be more seriously maintained and the results of well-conducted quasi-experiments considered as serious research.

The group concurred that the first essential standard for quasi-experiments should be adequate pretesting of participants. Invariably, several pretest measures are required to demonstrate comparability between groups. Moreover, these measures must have documented reliability and validity. Quasi-experiments with no pretest data should not be considered acceptable for publication in journals or for widespread dissemination unless strong disclaimers are provided. Also, studies in which there is more than a 0.5 *SD* difference in pretest scores on important variables should be carefully scrutinized before they are published in journals or used as evidence to support the effectiveness of a particular intervention or instructional approach.

Another issue that emerged in the discussions was the fragility of analysis of covariance as a means of correcting for initial pretest differences between experimental and comparison samples. Use of analysis of covariance should be limited to cases where initial differences are quite small—that is, no more than one-half *SD*. Furthermore, the standard assumptions mentioned in every textbook must be met.

Growth curve analysis can be used in quasi-experiments if more than two testing occasions are included in the design (see Lyon and Moats, 1997, for an introduction to this topic and its relevance for special education research; for a more detailed treatment of the topic, see e.g., Bryk & Raudenbush, 1992). With growth curve analysis, the likelihood of erroneous inferences is somewhat reduced. The reason for this is that, with growth curve analysis, it becomes clear when students are from different populations (or from populations that react differentially to the intervention). With analysis of co-

variance, the assumption is made that all students are from the same population. Unless this assumption is true, attempts at statistical control for initial differences are ineffective or invalid (Campbell & Erlebacher, 1970).

Increasingly, researchers are using both the student and the class as a unit of analysis. This multilevel approach to data analysis is optimal whenever sample size is large (i.e., at least 100 students per condition).

For all quasi-experiments, the authors should explain how they attempted to control for extraneous variables (such as confounding the intervention with teacher effectiveness), include a rationale for why their particular quasi-experimental design was employed, and provide some type of disclaimer. If the authors are candid about the design limitations, indicate that the results are largely exploratory, and link the findings to prior research, quasi-experiments can contribute to the accumulated knowledge on a topic. For quasi-experiments, the responsibility is on the researchers to demonstrate that the effects are due to the intervention rather than to other, extraneous factors. This burden of proof is much heavier than in studies with random assignment. A summary of the main points in this section is provided in Figure 2.

1. Provide a thorough description of samples.
2. Strive for random assignment of students to treatment conditions.
3. Explore the use of alternative designs, such as formative or design experiments.
4. Quasi-experiments need to be critically reviewed.
   a. Groups should not exceed more than 0.5 standard deviation (*SD*) units on salient pretest variables.
   b. Thorough sample description and analysis of comparison groups is essential.

FIGURE 2. Recommendations for probing the nature of the independent variable.

Selection of Dependent Measures

Far too often, the weakest part of an intervention study is the quality of the measures used to evaluate the impact of the intervention. Thus, a good deal of the researcher's effort should be devoted to selection and development of dependent measures. In essence, the conclusion of a study depends not only on the quality of the intervention and the nature of the comparison groups, but also on the quality of the measures selected or developed to evaluate intervention effects.

Intervention researchers often spend more time on aspects of the intervention related to instructional procedures than on dependent measures. Although it is understandable that many educators would rather teach than test, and few students enjoy being tested, creating tests of unknown validity invariably weakens the power of a study (i.e., reduces the chances of documenting the intervention's effectiveness) and limits the potential for synthesis because of a lack of common measures.

The importance of selecting, developing, and refining dependent measures to address the full array of research questions in a study is critical in high-quality research. It is usually valuable to use multiple measures in a study, given that any measure is necessarily incomplete and imperfect and no one measure can represent all, or even most, of the important phenomena that an intervention might affect. Any one measure can assess only a facet of the construct of interest, and, therefore, that measure is necessarily narrow or restricted. However, it is also important, when using multiple dependent measures, to ensure that appropriate statistical analyses are used to avoid inflating the possibility of finding significant effects.

Note that in many intervention studies researchers may assess more than one construct. For example, in a study examining the effects of socially mediated instruction on reading outcomes, a team may want to assess differences in terms of reading outcomes as well as social relations. For each construct, more than one measure will be needed for valid assessment. With reading comprehension, for example, it may be important to ascertain whether the experimental condition affects both literal and inferential comprehension. For social relations, it may be important to ascertain whether the effects are specific to general social standing or to close personal friendships.

Also, students may respond in a variety of ways as they are assessed in a particular area. For example, a test of reading comprehension may require students to (a) read a sentence, paragraph, or story silently or orally and write or say answers to multiple choice, short answer, or essay questions; (b) orally or silently read sentences or passages that contain blanks and restore those blanks with semantically correct words either orally or in writing; or (c) read paragraphs or stories and write or tell summaries of the content read. Although each type of measure taps some dimension of reading comprehension, it should be clear that the testing requirements differ dramatically, and student performance—and consequently assessed treatment effects—may vary accordingly.

Guidelines for Selecting Measures

Most of the constructs used in educational research—mathematical problem solving, expressive writing ability, phonemic awareness, self-esteem—are at best loosely defined. For this reason, multiple measures are always a necessity, as is a clear rationale of how a particular construct is being used within a study. It is not surprising that there is an art to selecting and developing measures.

Advantages of Using Multiple Measures

One of the foremost challenges to a research team is the selection of measures that are well aligned with the substance of the intervention and that will be sensitive to (i.e., register) treatment effects. The second and often competing challenge to researchers is the selection or development of measures that are sufficiently broad and robust to avoid criticism for "teaching to the test" through the specific intervention and to demonstrate that generalizable skills have been successfully taught. An intervention may have adverse effects, or additional benefits, that a researcher should attempt to identify by using measures that are sensitive to other potential effects. For example, exposure to well-designed instruction in solving fraction problems may lead to an enhanced conceptual understanding of fractions. Similarly, a well-implemented conceptual approach for teaching students with disabilities about fractions may, as an additional effect, lead to enhanced ability in computation and word problem solving. Only by using a broad array of measures can we begin to understand these influences empirically. Finally, the benefits of an intervention are clearer if its effects converge across multiple sources.

Overreliance on Closely Aligned Measures

By and large, intervention effects are stronger the more closely aligned measures are to the specific objectives of the intervention. There are several good reasons for including measures closely aligned to the effects of the intervention. One reason is that it makes sense to measure the extent to which students learn precisely what they are taught. For example, can students who are taught to use story grammar questions actually answer story grammar questions? Can students who recently completed a unit on addition and subtraction of fractions actually add and subtract fractions? Can students who participated in a program to increase their reading fluency actually read more fluently?

Although one component of a measurement battery may be biased in favor of the experimental intervention, the data on that measure still provide critical information. Being clear about this bias also is important, of course. It should become increasingly standard practice to provide the data and to include appropriate caveats.

For example, L. S. Fuchs and Fuchs (1994) described data from one of their studies involving teacher use of story grammar questions. They noted that the use of story grammar questions did not have an effect on students' oral reading fluency. But for measures that were more closely related to the intervention, such as the Stanford Achievement Test, which requires students to read passages and respond to multiple choice questions, the effect size was a moderate .37 favoring those students whose teachers incorporated more story grammar comprehension questions into instruction. For the retell task—the measure most closely aligned to the intervention—the effect size was strongest, .67.

Other researchers have stressed the importance of selecting an array of measures that are not heavily biased toward the intervention. Usually, these are measures that the researchers have not developed themselves. Swanson and Hoskyn (1998) noted that treatment effects were stronger on experimenter-developed measures than on standardized measures of the same construct. For example, if researchers are investigating story grammar, their assessment is quite likely to include several measures of story grammar knowledge and use. Other broad measures of comprehension also should be used, and the effect that the intervention has on these measures almost invariably will be less pronounced than on the more closely aligned story grammar questions. In the view of many researchers, the ultimate goal is to build broad competencies in students, and, too often, limited experimenter-developed measures tap very small aspects of these abilities. In other words, when studies show clear effects on broad-spectrum tests (e.g., a well-standardized achievement measure) without having explicitly taught the content of the test, we should place substantial weight on the importance of the findings.

The important guideline is to use a *combination* of broad measures and measures closely aligned to the intervention. When researchers exclusively use off-the-shelf, commonly available measures, such as published standardized achievement tests or well-known self-concept measures such as the Harter scales (Harter, 1985), the data may be insensitive to the effects of the intervention. Conversely, when researchers rely only on researcher-developed measures, it is unclear to what extent the outcomes are generalizable to critical constructs of interest. Also, experimenter-developed measures frequently do not meet acceptable standards of reliability and validity.

It would benefit the field to make clear the distinctions between experimenter-developed measures and those measures that are familiar to readers, such as published norm-referenced tests or commonly used measures of social behavior. Such distinctions should be made both empirically and conceptually. By including well-known assessments to supplement those developed by the researcher, the potential for bias is reduced. Presenting intercorrelations between measures gives readers a sense of the extent of overlap and concurrent validity of the measures. Researchers should include discussions of construct validity, using their own data and data from other relevant research.

The fate of any study is in large part due to the quality of the measures developed and selected. We strongly recommend that pilot research be used to refine measures and that piloting procedures and psychometric characteristics be reported for all measures used. There is a tendency to use the best passages (in reading research) and the best problems (in math or science research) for the teaching and save the weaker ones for testing. After all, we want the intervention to be the best possible, and we want to use the best materials available to obtain the desired effects. Yet this pattern is unwise for

group research. In essence, the findings in any intervention study depend on the quality of the individual items selected for assessment. Consequently, using weaker items (passages or problems), or items of unknown quality, for assessing outcomes while saving the better ones for instruction is ill advised.

Although conventional psychometric indices have their limitations, we believe they often are underutilized in special education research. In particular, the estimation of internal consistency reliability (often referred to by technical names such as coefficient alpha or Cronbach's alpha) is a critical aspect of a research study. This procedure is often neglected, however, even in published research articles. This omission is difficult to understand, because coefficient alpha is easy to compute with common statistical packages and the information is very important. Internal consistency reliability programs help us understand how well a cluster of items on a test fit together—how well performance on one item predicts performance on another. These analyses help researchers locate items that don't fit well—that is, items with a weak item-to-total correlation. Revising these items, or dropping them, can improve statistical power and, ultimately, the quality of the study.

Research teams also may want to test new or labor-intensive measures, such as think-alouds and retellings, on a small random subsample of students in the study. This is an effective way to experiment with innovative and potentially powerful assessment methods. The findings, although exploratory in nature, can contribute possible interesting insights in the studies and help inform future research. In Figure 3, we present a summary of issues to consider in the development and use of dependent measures for group research.

1. Select some measures that are not aligned tightly to the intervention.
2. Ensure that not all measures are experimenter developed and that some have been validated in prior research; and assess broad aspects of the construct being investigated.
3. Calculate coefficient alpha for new measures and observations.
4. Seek a balance between global and specific measures.
5. Look at intervention research as an opportunity to really build understanding of measures—contrast validity.

FIGURE 3. Recommendations regarding the use of dependent measures.

The Importance of Replications

In 1978, Gage responded to the growing numbers of researchers and journalists who suggested that no useful, generalizable knowledge had come, or would emerge, from educational research that could improve teaching or learning. Gage predicted that data would emerge over time to support empirical studies of teaching and learning using quantitative methods. By 1997, Gage's prediction had "been resoundingly upheld by the hundreds of meta-analyses that have been reported in the intervening years" (Gage, 1997, p. 19). Gage's (1997) point was that "many generalizations in education do hold up across many replications with *high consistency* [italics added]. . . despite the fact that replications inevitably differ in the persons studied, in the measurement methods used, in the social contexts involved, and in other ways" (p. 19).

Research in special education has contributed substantially to the knowledge base on effective educational practices. Numerous meta-analyses (Elbaum et al., 1999; Scruggs & Mastropieri, 1994; Swanson & Hoskyn, 1998) have been conducted and have provided confirmations of findings from seminal studies. Replications have converged to form a consistent knowledge base that generalizes across student, teacher, and environmental variables (see Forness, Kavale, Blum, & Lloyd, 1997). Several aspects of replication studies seem especially important in special education, and two in particular seem worth discussing here.

Independent Replications

First, replication by researchers not invested in developing the independent variable should be a standard practice of special education research. Walberg and Greenberg (1998) pointed out that a major failure in contemporary educational research is the small number of independent replications. In special education, there is a small but growing tradition of independent replication studies. Hasselbring, Sherwood, and Bransford (1986), for example, investigated the effectiveness of video disc programs they had no role in developing. Similarly, Klingner and Vaughn (1996) investigated the effects of reciprocal teaching, and Simmons, Fuchs, Fuchs, Hodge, and Mathes (1994) studied the conditions under which the effects of peer tutoring were enhanced. Walberg and Greenberg (1998) discussed the importance of independent replication with extremely popular programs, such as *Success for All*. They pointed out that evaluations of *Success for All* carried out by the developers of the program resulted in numerical effects on achievement that were among the largest reported in the literature. Independent evaluations tended to result in much more modest effects.

Replications Employing Components Analysis

A second standard replication practice in special education should be conducting studies to determine which components of a complex instructional approach are critical for achieving an impact on learning or social competence. In a study using a contrasted groups design, Gersten et al. (1982) observed in-

structional interactions of 11 teachers implementing an approach to teaching beginning reading to at-risk students. The two variables that tended to most clearly distinguish the best teachers from the poorest teachers were (a) responding to student errors and problems immediately, and (b) maintaining a success rate of at least 85% with all students, even those placed in the lowest reading group. These findings were replicated with a larger sample of teachers the following year (Gersten et al., 1986). Stallings (1980), Leinhardt, Zigmond, and Cooley (1981), and Brophy and Good (1986) also documented the importance of these two instructional variables—providing immediate feedback when students make oral reading errors and having students read with high levels of success. Unlike replication studies where impartial independent investigation is optimal, these components analysis studies warrant investigation by researchers with an in-depth understanding of the intervention.

Next Steps

In the 5 years of meetings supported by the Office of Special Education Programs and the Council for Exceptional Children, two areas of broad consensus emerged for improving the quality of research and our ability to synthesize research. The first was that researchers should routinely report effect sizes as well as probability values from statistical tests. An excellent resource for a conceptual treatment of quantitative research syntheses and for procedures for the calculation of effect sizes is Cooper and Hedges (1994).

Standard reporting formats should allow researchers to read a given study and extract the necessary information for classifying critical variables regarding the nature of the study and the participants. The study should also provide the necessary information to calculate effect sizes for each relevant dependent measure. Conveying information for the calculation of effect sizes is usually done best in tables displaying descriptive data, which normally include all relevant pretest and posttest mean scores and their standard deviations. Although the most frequent meta-analytic syntheses use effect size calculations on posttest scores, it is also acceptable to calculate posttest effect sizes after adjusting for pretest differences, especially when quasi-experimental designs are used. For this reason, it is important to also report pretest measures and standard deviations for experimental and comparison groups.

A second area of agreement was the need for researchers to use a set of common measures when researching a given topic. These measures would be selected by a group of researchers with expertise in that field. They would typically include standardized measures but might also include innovative measures that have been successfully used in research. They need not be limited to paper and pencil tests and could include, for example, direct observation techniques, measures of oral reading fluency following standardized procedures, theoretically compelling measures such as rapid automatized naming (Denckla & Rudel, 1976), or performance measures using state-of-the art scoring methodology.

This would not preclude the use of additional experimenter-developed measures to study specific aspects of an intervention. But routine use of a small set of core measures in areas such as oral reading, social competence, reading comprehension, and task persistence would enhance the process of research synthesis and integration of findings.

Our hope is to provide guidance for special education researchers about how to conduct high-quality experimental studies using group-contrast methods. To that end, we consulted with literally dozens of colleagues and identified several critical ways in which research teams can conduct high-quality studies. We do not offer these recommendations as an exhaustive set of guidelines but, rather, as suggestions for those who wish to use group-contrast designs and are predisposed to pursuing methods that enhance the quality of findings.

Our recommendations center on familiar issues—selecting research questions and conceptualizing their relation to other studies, assigning participants to groups, choosing and describing measures, describing participants and conditions, and so forth. Obviously, researchers who are interested can learn more about these issues by consulting standard textbooks on research design. However, we have presented these concepts with particular reference to special education intervention research and to how those issues present special challenges to those of us engaged in studying ways to provide better services to students with disabilities, their teachers, and their parents.

AUTHORS' NOTES

1. The preparation of this document was funded by the U.S. Department of Education, Office of Special Education Programs (OSEP), contract no. RR93002005 with The Council for Exceptional Children for the operation of the ERIC/OSEP special project through the ERIC Clearinghouse on Disabilities and Gifted Education. However, the views expressed here do not necessarily reflect the positions of OSEP or of the Department of Education.
2. The authors wish to thank the contributions of the following individuals involved in the project: Joanna Williams, Sharon Vaughn, Deborah Simmons, Lynn Fuchs, Lee Swanson, Susan Osborne, Martha Thurlow, Tom Keating, Shanna Hagan Burke, Douglas Carnine, and Sylvia Smith.

REFERENCES

Ball, D. L. (1990). Reflections and deflections of policy: The case of Carol Turner. *Educational Evaluation and Policy Analysis, 12,* 247–259.

Ball, D. L. (1995). Blurring the boundaries of research and practice. *Remedial and Special Education, 16,* 354–364.

Beck, I. L., McKeown, M. G., Sandora, C., Kucan, L., & Worthy, J. (1996). Questioning the author: A yearlong classroom implementation to engage students with text. *Elementary School Journal, 96,* 385–414.

Billups, L. H. (1997). Response to bridging the research-to-practice gap. *Exceptional Children, 63,* 525–526.

Bottge, B., & Hasselbring, T. S. (1993). A comparison of two approaches for teaching complex, authentic mathematics problems to adolescents in remedial math classes. *Exceptional Children, 59,* 556–566.

Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. Witrock (Ed.), *The third handbook of research on teaching* (pp. 328–375). New York: MacMillan.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of Learning Sciences, 2*(2), 141–178.

Brown, R., Pressley, M., Van Meter, P., & Schuder, T. (1996). A quasi-experimental validation of transactional strategies instruction with low-achieving second grade readers. *Journal of Educational Psychology, 88,* 18–37.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models.* Newbury Park, CA: Sage.

Campbell, D. T., & Erlebacher, A. E. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), *Compensatory education: A national debate* (Vol. 3, pp. 185–233). New York: Brunner/Mazel.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand-McNally.

Carnine, D. (1995). The professional context for collaboration and collaborative research. *Remedial and Special Education, 16,* 368–371.

Carnine, D., Jones, E., & Dixon, R. (1994). Mathematics: Educational tools for diverse learners. *School Psychology Review, 23,* 406–427.

Chamot, A. U., Keatley, C., & Mazur, A. (1999, April). *Literacy development in adolescent English language learners: Project Accelerated Literacy* (PAL). Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Charters, W. W., Jr., & Jones, J. E. (1974, February). *On neglect of the independent variable in program evaluation.* Eugene: University of Oregon, Project MITT.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand-McNally.

Cooley, W. W., & Leinhardt, G. (1980). The instructional dimensions study. *Education Evaluation and Policy Analysis, 2,* 7–25.

Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

Denckla, M. B., & Rudel, R. (1976). Rapid 'automatized' naming (RAN): Dyslexia differentiated from other learning disabilities. *Neuropsychologia, 14,* 471–479.

Dimino, J., Gersten, R., Carnine, D., & Blake, G. (1990). Story grammar: An approach for promoting at-risk secondary students' comprehension of literature. *Elementary School Journal, 91,* 19–32.

Dunst, C. J., & Trivette, C. M. (1994). Methodological considerations and strategies for studying the long-term effects of early intervention. In S. L. Friedman & H. C. Haywood (Eds.), *Developmental follow-up: Concepts, domains, and methods.* New York: Academic Press.

Echevarria, J. (1995). Interactive reading instruction: A comparison of proximal and distal effects of instructional conversations. *Exceptional Children, 61,* 536–552.

Elbaum, B., Vaughn, S., Hughes, M., & Moody, S. W. (1999). Grouping practices and reading outcomes for students with disabilities. *Exceptional Children, 65,* 399–415.

Englert, C. S., Raphael, T. E., Anderson, L. M., Anthony, H. M., & Stevens, D. D. (1991). Making writing strategies and self-talk visible: Cognitive strategy instruction in regular and special education classrooms. *American Educational Research Journal, 28,* 337–372.

Englert, C. S., & Tarrant, K. L. (1995). Creating collaborative cultures for educational change. *Elementary School Journal, 16,* 325–336.

Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology, 90*(1), 37–55.

Forness, S., & Kavale, K. A. (1987). De-psychologizing special education. In R. B. Rutherford & C. M. Nelson (Eds.), *Severe behavior disorders of children and youth* (pp. 2–24). Boston: College-Hill.

Forness, S., Kavale, K. A., Blum, I. M., & Lloyd, J. W. (1997). Mega-analysis of meta-analysis. *Teaching Exceptional Children, 29*(6), 4–9.

Fuchs, D., & Fuchs, L. S. (1998). Researchers and teachers working together to adapt instruction for diverse learners. *Learning Disabilities Research & Practice, 13,* 126–137.

Fuchs, D., Fuchs, L. S., Mathes, P. G., & Lipsey, M. W. (in press). Reading differences between underachievers with and without learning disabilities: A meta-analysis. In R. Gersten, E. Schiller, S. Vaughn, & J. Schumm (Eds.), *Research syntheses in special education.* Mahwah, NJ: Erlbaum.

Fuchs, D., Fuchs, L. S., Mathes, P. H., & Simmons, D. C. (1997). Peer-assisted strategies: Making classrooms more responsive to diversity. *American Educational Research Journal, 34,* 174–206.

Fuchs, L. S., & Fuchs, D. (1994). Academic assessment and instrumentation. In S. Vaughn & C. Bos (Eds.), *Research issues in learning disabilities: Theory, methodology, assessment, and ethics* (pp. 233–242). New York: Springer Verlag.

Fuchs, L. S., Fuchs, D., Bentz, J., Phillips, N. B., & Hamlett, C. L. (1994). The nature of student interactions during peer tutoring with and without prior training and experience. *American Educational Research Journal, 31,* 75–103.

Fuchs, L. S., Fuchs, D., Kazdan, S. R., & Allen, S. (1999). Effects of peer-assisted learning strategies in reading with and without training in elaborated help giving. *Elementary School Journal, 99,* 201–220.

Gage, N. L. (1978). *The scientific basis of the art of teaching.* New York: Teachers College Press.

Gage, N. L. (1997). "The vision thing": Educational research and AERA in the 21st century part I: Competing visions of what educational researchers should do. *Educational Researcher, 26*(4), 18–21.

Gall, M., Borg, W., & Gall, J. (1996). *Educational research: An introduction* (6th ed.). White Plains, NY: Longman.

Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. (1982). *Stanford achievement test.* San Antonio, TX: Psychological Corp.

Gersten, R. (1996). Literacy instruction for language-minority students: The transition years. *Elementary School Journal, 96,* 227–244.

Gersten, R., & Baker, S. (1997). *Design experiments, formative experiments, and developmental studies: The changing race of instructional research in special education.* (Technical Report 98-1). Eugene, OR: Eugene Research Institute.

Gersten, R., & Baker, S. (in press). What we know about effective instructional practices for English-language learners. *Exceptional Children.*

Gersten, R., Carnine, D., & Williams, P. (1982). Measuring implementation of a structured educational model in an urban setting: An observational approach. *Educational Evaluation and Policy Analysis, 4,* 67–79.

Gersten, R., Carnine, D., Zoref, L., & Cronin, D. (1986). A multifaceted study of change in seven inner city schools. *Elementary School Journal, 86,* 257–276.

Gersten, R., & McInerney, M. (1997). *What parents, teachers, and researchers view as critical issues in special education research* (Technical Report 97-1). Washington, DC: American Institutes for Research.

Gersten, R., Vaughn, S., Deshler, D., & Schiller, E. (1997). What we know about using research findings: Implications for improving special education practice. *Journal of Learning Disabilities, 30,* 466–476.

Gersten, R., Williams, J., Fuchs, L., Baker, S., Koppenhaver, D., Spadorcia, S., & Harrison, M. (1998). *Improving reading comprehension for children with disabilities: A review of research.* Washington, DC: American Institutes for Research.

Good, T. L., & Grouws, D. A. (1977). Teaching effects: A process product study in fourth grade mathematics classrooms. *Journal of Teacher Education, 28,* 49–54.

Graham, S., & Harris, K. R. (1989). Components analysis of cognitive strategy instruction: Effects on learning disabled students' compositions and self-efficacy. *Journal of Educational Psychology, 81,* 353–361.

Graham, S., & Harris, K. (1994). Cognitive strategy instruction: Methodological issues and guidelines in conducting research. In S. Vaughn & C. Bos (Eds.), *Research issues in learning disabilities: Theory, method-*

ology, assessment, and ethics (pp. 146–160). New York: Springer Verlag.

Graham, S., MacArthur, C., & Schwartz, S. (1995). Effects of goal setting and procedural facilitation on the revising behavior and writing performance of students with writing and learning problems. *Journal of Educational Psychology, 87,* 230–240.

Greeno, J. G., & Middle School Mathematics Through Applications Project Group. (1998). The situativity of knowing, learning, and research. *American Psychologist, 53,* 5–26.

Greenwood, C. R., & Delquadri, J. (1988). Code for instructional structure and student academic response (CISSAR). In M. Hersen & A. S. Bellack (Eds.), *Dictionary of behavioral assessment* (pp. 120–122). New York: Pergamon.

Gresham, F. M., Gansle, K. A., Noell, G. H., Cohen, S., & Rosenblum, S. (1993). Treatment integrity of school-based behavioral intervention studies: 1980–1990. *School Psychology Review, 22,* 254–272.

Hall, G. E., & Loucks, S. (1977). A developmental model for determining whether the treatment is actually implemented. *American Educational Research Journal, 14,* 264–276.

Harter, S. (1985). *The self-perception profile for children.* Denver, CO: University of Denver.

Hasselbring, T., Sherwood, B., & Bransford, J. (1986). *An evaluation of the Mastering Fractions level-one instructional videodisc program.* Nashville, TN: George Peabody College of Vanderbilt University, The Learning Technology Center.

Hasselbring, T., Sherwood, R., Bransford, J., Fleenor, K., Griffith, D., & Goin, L. (1988). Evaluation of a level-one instructional videodisc program. *Journal of Educational Technology Systems, 16,* 151–169.

Hunt, P., & Goetz, L. (1997). Research on inclusive educational programs, practices, and outcomes for students with severe disabilities. *The Journal of Special Education, 31,* 3–29.

Kelly, B., Gersten, R., & Carnine, D. (1990). Student error patterns as a function of curriculum design. *Journal of Learning Disabilities, 23,* 23–32.

Kennedy, M. M. (1991). Implications for teaching. In E. A. Ramp & C. S. Pederson (Eds.), *Follow through: Program and policy issues* (pp. 57–71). Washington, DC: U.S. Department of Education, Office of Education Research and Improvement.

Kennedy, M. M. (1997). The connection between research and practice. *Educational Researcher, 26*(7), 4–12.

Kline, F. M., Deshler, D. D., & Schumaker, J. B. (1992). Implementing learning strategy instruction in class settings: A research perspective. In M. Pressley, K. R. Harris, & J. T. Guthrie (Eds.), *Promoting academic competence and literacy in school* (pp. 361–406). San Diego: Academic Press.

Klingner, J. K., & Vaughn, S. (1996). Reciprocal teaching of reading comprehension strategies for students with learning disabilities who use English as a second language. *Elementary School Journal, 96,* 275–293.

Klingner, J. K., Vaughn, S., & Schumm, J. S. (1998). Collaborative strategic reading during social studies in heterogeneous fourth-grade classrooms. *Elementary School Journal, 99,* 3–22.

Kornblet, A. (1997). Response to bridging the research-to-practice gap. *Exceptional Children, 63,* 523–524.

Krathwohl, D. R. (1993). *Methods of educational and social science research.* White Plains, NY: Longman.

Leinhardt, G. (1977). Program evaluation: An empirical study of individualized instruction. *American Educational Research Journal, 14,* 277–293.

Leinhardt, G., Zigmond, N., & Cooley, W. (1981). Reading instruction and its effects. *American Educational Research Journal, 18,* 343–361.

Losardo, A., & Bricker, D. (1994). A comparison study: Activity-based intervention and direct instruction. *American Journal on Mental Retardation, 98,* 744–765.

Lovett, M. H., Borden, S. H., DeLuca, T., Lacerenza, L., Benson, N. J., & Brackstone, D. (1994). Treating the core deficits of developmental dyslexia: Evidence of transfer of learning after phonologically and strategy based reading training programs. *Developmental Psychology, 30,* 805–822.

Lyon, G. R., & Moats, L. C. (1997). Critical conceptual and methodological considerations in reading intervention research. *Journal of Learning Disabilities, 30,* 578–588.

Malouf, D. B., & Schiller, E. P. (1995). Practice and research in special education. *Exceptional Children, 61,* 414–424.

McKinney, J. D., Osborne, S. S., & Schulte, A. C. (1993). Academic consequences of learning disability: Longitudinal prediction of outcomes at 11 years of age. *Learning Disabilities Research & Practice, 8,* 19–27.

National Center to Improve the Tools of Educators. (1998). *Evaluation of research on educational approaches (EREA).* Unpublished manuscript, University of Oregon.

Newman, D. (1990). Opportunities for research on the organizational impact of school computers. *Educational Researcher, 19,* 8–13.

O'Connor, R. E., & Jenkins, J. R. (1995). Improving the generalization of sound/symbol knowledge: Teaching spelling to kindergarten children with disabilities. *The Journal of Special Education, 29,* 255–275.

O'Connor, R., Jenkins, J., Leicester, N., & Slocum, T. (1993). Teaching phonological awareness to young children with learning disabilities. *Exceptional Children, 59,* 532–546.

O'Connor, R. E., Jenkins, J. R., & Slocum, T. A. (1995). Transfer among phonological tasks in kindergarten: Essential instructional content. *Journal of Educational Psychology, 37*(2), 202–217.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1,* 117–175.

Peterson, L., Homer, A. L., & Wonderlich, S. A. (1982). The integrity of independent variables in behavior analysis. *Journal of Applied Behavior Analysis, 15,* 477–492.

Pressley, M., & El-Dinary, P. B. (1997). What we know about translating comprehension-strategies instruction research into practice. *Journal of Learning Disabilities, 30,* 486–488.

Pressley, M., & Harris, K. R. (1994). Increasing the quality of educational intervention research. *Educational Psychology Review, 6,* 191–208.

Reinking, D., & Pickle, J. M. (1993). Using a formative experiment to study how computers affect reading and writing in classrooms. In C. Z. Kinzer & D. J. Leu (Eds.), *Examining central issues in literacy research, theory, and practice* (pp. 263–270). Chicago: National Reading Conference.

Richardson, V. (1994). Conducting research on practice. *Educational Researcher, 23*(5), 5–10.

Richardson, V., & Anders, P. (1998). A view from across the Grand Canyon. *Learning Disability Quarterly, 21,* 85–97.

Rosenberg, M. S., Bott, D., Majsterek, D., Chiang, B., Simmons, D., Gartland, D., Wesson, C., Fraham, S., Smith-Myles, B., Miller, M., Swanson, H. L., Bender, W., Rivera, D., & Wilson, R. (1994). Minimum standards for the description of participants in learning disabilities research. *Remedial and Special Education, 15*(1), 56–59.

Rosenshine, B., & Meister, C. (1994). Reciprocal teaching: A review of the research. *Review of Educational Research, 64,* 479–530.

Sawyer, R. J., Graham, S., & Harris, K. R. (1992). Direct teaching, strategy instruction, and strategy instruction with explicit self-regulation: Effects on the composition skills and self-efficacy of students with learning disabilities. *Journal of Educational Psychology, 84,* 340–352.

Scanlon, D., Schumaker, J. B., & Deshler, D. D. (1994). Collaborative dialogues between teachers and researchers to create education interventions: A case study. *Journal of Educational and Psychological Consultation, 5*(1), 69–76.

Scruggs, T. E., & Mastropieri, M. A. (1994). Issues in conducting intervention research: Secondary students. In S. Vaughn & C. Bos (Eds.), *Research issues in learning disabilities: Theory, methodology, assessment, and ethics* (pp. 130–145). New York: Springer Verlag.

Simmons, D. C., Fuchs, D., Fuchs, L. S., Hodge, J. P., & Mathes, P. G. (1994). Importance of instructional complexity and role reciprocity to classwide peer tutoring. *Learning Disabilities Research & Practice, 9,* 203–212.

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

Slavin, R., & Madden, N. (1995, April). *Effects of Success for All on the achievement of English language learners.* Paper presented at the annual meeting for the American Educational Research Association, San Francisco.

Slocum, T. A., O'Connor, R. E., & Jenkins, J. R. (1993). Transfer among phonological manipulation skills. *Journal of Educational Psychology, 85,* 618–630.

Stallings, J. (1975). *Follow through program classroom observation evaluation.* Menlo Park, CA: Stanford Research Institute.

Stallings, J. (1980). Allocated academic learning time revisited, or beyond time on task. *Educational Leadership, 9*(11), 11–16.

Swanson, H. L., & Hoskyn, M. (1998). Experimental intervention research on students with learning disabilities: A meta-analysis of treatment outcomes. *Review of Educational Research, 68,* 277–321.

Troia, G. A. (1999). Phonological awareness intervention research: A critical review of the experimental methodology. *Reading Research Quarterly, 34,* 28–52.

Vockell, E. L., & Asher, J. W. (Eds.). (1995). *Educational research* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Walberg, H. J., & Greenberg, R. C. (1998, April 8). The Diogenes factor: Why it's hard to get an unbiased view of programs like Success for All. *Education Weekly,* p. 52.

Williams, J. P., Brown, L. G., Silverstein, A. K., & deCani, J. S. (1994). An instructional program in comprehension of narrative themes for adolescents with learning disabilities. *Learning Disability Quarterly, 17,* 205–221.

Williams, S. R., & Baxter, J. A. (1996). Dilemmas of discourse-oriented teaching in one middle school mathematics classroom. *Elementary School Journal, 97,* 21–38.

Wong, B. Y. L., Butler, D. L., Ficzere, S. A., & Kuperis, S. (1997). Teaching adolescents with learning disabilities and low achievers to plan, write, and revise compare–contrast essays. *Learning Disabilities Research & Practice, 12,* 2–15.

Woodward, J., & Gersten, R. (1992). Innovative technology for secondary learning disabled students: A multi-faceted study of implementation. *Exceptional Children, 58,* 407–421.

Ysseldyke, J. E., & Christenson, S. L. (1987). *The instructional environment scale: A comprehensive methodology for assessing an individual student's instruction.* Austin, TX: PRO-ED.