

Quality Indicators for Group Experimental and Quasi-Experimental Research in Special Education

RUSSELL GERSTEN
Instructional Research Group

LYNN S. FUCHS
DONALD COMPTON
Vanderbilt University

MICHAEL COYNE
University of Connecticut

CHARLES GREENWOOD
University of Kansas

MARK S. INNOCENTI
Utah State University

ABSTRACT: *This article presents quality indicators for experimental and quasi-experimental studies for special education. These indicators are intended not only to evaluate the merits of a completed research report or article but also to serve as an organizer of critical issues for consideration in research. We believe these indicators can be used widely, from assisting in the development of research plans to evaluating proposals. In this article, the framework and rationale is explained by providing brief descriptions of each indicator. Finally, we suggest a standard for determining whether a practice may be considered evidence-based. It is our intent that this standard for evidenced-based practice and the indicators be reviewed, revised as needed, and adopted by the field of special education.*

This article presents a set of quality indicators for experimental and quasi-experimental studies for the field of special education. We believe there is a critical need for

such a set of indicators, given the current focus on the need for an increase in rigorous, scientific research in education. Recently, the National Research Council (NRC, 2002), in a report on scientific research in education, noted that they saw

no reason why education could not be subject to the same scientific methods as other disciplines such as chemistry or physics. They further argued that there is a place for all research methodologies in educational research: survey research, qualitative research, and correlational research. As Feuer, Towne, and Shavelson (2002) noted, "If a research conjecture or hypothesis can withstand scrutiny by multiple methods, its credibility is enhanced greatly. Overzealous adherence to the use of any given research design flies in the face of this fundamental principle" (p. 8).

Yet the NRC report is unequivocal in stating numerous times that experiments using randomized trials are currently underutilized in educational research, despite being "the single best methodological route to ferreting out systematic relations between actions and outcome" (Feuer et al., 2002, p. 8). In that sense, they reiterate a point made over a generation ago by Campbell and Stanley (1966), who highlighted that controlled experimental research conducted in field settings was

the...only means for settling disputes regarding educational practice, as the only way of verifying educational improvements, and as the only way of establishing a cumulative tradition in which improvement can be introduced without the danger of a faddish discard of old wisdom in favor of inferior novelties. (p. 2)

Until recently, there was no specific set of standards or quality indicators for evaluating the quality of either proposed or completed experimental or quasi-experimental intervention research. Few of the numerous educational research textbooks provide a clear list of criteria that fit the realities of contemporary field research, although they present excellent ideas (e.g., Boruch, 1997; Gall, Borg, & Gall, 2002; Shadish, Cook, & Campbell, 2002). Fortuitously, slightly before we began this endeavor, the What Works Clearinghouse of the U.S. Department of Education released the *Study Design and Implementation Assessment Device (DIAD)* (Valentine & Cooper, 2003). Its major goal is to evaluate whether a research article or research report can be considered sufficiently valid and reliable to be entered into a research synthesis on the effectiveness of an intervention or approach.

Our goal is a bit broader. We intend these quality indicators to be used not only to evaluate the merits of a completed research report or article, but also to evaluate research proposals, dissertation proposals, and grant applications submitted to funding agencies. We also intend this to be useful to researchers as they think through the design of a study, to serve as a checklist or organizer of critical issues for consideration.

We intentionally chose to look at quality indicators, rather than a set of "yes/no" standards, because evaluating a research design always entails weighing the relative strengths and weaknesses in a set of domains. Our goal is not to provide a basic primer on how to design a high-quality field research study. There are numerous introductory (e.g., Gall et al., 2002) and advanced texts (e.g., Shadish et al., 2002) widely available to explain the many concepts underlying the design of experimental studies. Our hope is that our set of indicators will be field-tested and refined, and then considered useful by journal editors and reviewers of federal grant proposals. We also envision this set of indicators assisting researchers as they design studies and guiding practitioners as they consider alternative educational practices for adoption in their classrooms and schools.

We intend these quality indicators to be used not only to evaluate the merits of a completed research report or article, but also to evaluate research proposals, dissertation proposals, and grant applications submitted to funding agencies.

Quality indicators for group experimental and quasi-experimental research proposals in special education are presented in Table 1. Table 2 presents a set of quality indicators for evaluating the quality of completed experimental or quasi-experimental studies. These tables address similar issues, but serve slightly different purposes. Both tables make a distinction between indicators that are considered *essential* for quality versus indicators that are considered to be *desirable* to have in a study.

TABLE 1*Essential and Desirable Quality Indicators for Group Experimental and Quasi-Experimental Research Proposals*

Essential Quality Indicators*Conceptualization Underlying the Study*

1. Is a compelling case for the importance of the research made? Is the conceptualization based on well-designed studies and does it reflect the scope of extant knowledge?
2. If an innovative approach is proposed, is it based on a sound conceptualization formed from sound research?
3. Are the research questions appropriate and stated clearly for the purposes of this study? Are valid arguments supporting the nature of intervention in the comparison group(s) presented?

Participants/Sampling

1. Will appropriate procedures be used to ensure that participants are comparable across intervention conditions on relevant characteristics? If random assignment is used, will information about participants prior to the intervention be made available to ensure that samples are comparable on salient characteristics?
2. Will sufficient information be provided to determine (or confirm) whether the participants demonstrated the disability(ies) or learning/social learning difficulties presented?
3. Will appropriate procedures be used to increase the probability that teachers or interventionists will be comparable across conditions?

Implementation of the Intervention and the Nature of Comparison Condition(s)

1. Is the intervention clearly described?
2. Are procedures for ensuring and assessing fidelity of implementation described?
3. Is the nature of instruction/services provided in comparison conditions described?

Outcome Measures

1. Will multiple measures be used to provide an appropriate balance between measures closely aligned with the intervention and measures of generalized performance?
2. Is evidence of reliability for the outcome measures provided? If not, will it be calculated?

Quality Indicators for Data Analysis

1. Are the data analysis techniques appropriate and linked to key research questions and hypotheses?
2. Is the variability within each sample accounted for either by sampling techniques (e.g., restricting range) or statistical techniques (blocking, analysis of covariance, growth curve analysis)?
3. Is a power analysis provided to describe the adequacy of the minimum cell size?

Desirable Quality Indicators

1. Are data collectors and/or scorers blind to study conditions and equally (un)familiar to examinees across study conditions?
 2. Does the study provide not only internal consistency reliability but also test-retest reliability and interrater reliability (when appropriate) for outcome measures?
 3. Are outcomes for capturing the intervention's effect measured beyond an immediate posttest?
 4. Is evidence of the validity of the measures provided? If not, will it be estimated based on data from the proposed study or with data collected from other samples?
 5. Will the research team assess more than surface features of fidelity implementation (e.g., number of minutes allocated to the intervention or teacher/interventionist following procedures specified)? Additionally, will the research team examine the quality of implementation?
 6. Will the research include actual audio or videotape excerpts that capture the nature of the intervention?
 7. Does the researcher conduct power analyses proper for varying levels of statistical analysis? For example, if data will be analyzed at a classroom or day care center level, are analyses at that level sensitive enough to detect effects?
-

TABLE 2*Essential and Desirable Quality Indicators for Group Experimental and Quasi-Experimental Research Articles and Reports***Essential Quality Indicators***Quality Indicators for Describing Participants*

1. Was sufficient information provided to determine/confirm whether the participants demonstrated the disability(ies) or difficulties presented?
2. Were appropriate procedures used to increase the likelihood that relevant characteristics of participants in the sample were comparable across conditions?
3. Was sufficient information given characterizing the interventionists or teachers provided? Did it indicate whether they were comparable across conditions?

Quality Indicators for Implementation of the Intervention and Description of Comparison Conditions

1. Was the intervention clearly described and specified?
2. Was the fidelity of implementation described and assessed?
3. Was the nature of services provided in comparison conditions described?

Quality Indicators for Outcome Measures

1. Were multiple measures used to provide an appropriate balance between measures closely aligned with the intervention^a and measures of generalized performance?
2. Were outcomes for capturing the intervention's effect measured at the appropriate times?

Quality Indicators for Data Analysis

1. Were the data analysis techniques appropriately linked to key research questions and hypotheses? Were they appropriately linked to the unit of analysis in the study?
2. Did the research report include not only inferential statistics but also effect size calculations?

Desirable Quality Indicators

1. Was data available on attrition rates among intervention samples? Was severe overall attrition documented? If so, is attrition comparable across samples? Is overall attrition less than 30%?
2. Did the study provide not only internal consistency reliability but also test–retest reliability and interrater reliability (when appropriate) for outcome measures? Were data collectors and/or scorers blind to study conditions and equally (un)familiar to examinees across study conditions?
3. Were outcomes for capturing the intervention's effect measured beyond an immediate posttest?
4. Was evidence of the criterion-related validity and construct validity of the measures provided?
5. Did the research team assess not only surface features of fidelity implementation (e.g., number of minutes allocated to the intervention or teacher/interventionist following procedures specified), but also examine quality of implementation?
6. Was any documentation of the nature of instruction or series provided in comparison conditions?
7. Did the research report include actual audio or videotape excerpts that capture the nature of the intervention?
8. Were results presented in a clear, coherent fashion?

^aA study would be acceptable if it included only measures of generalized performance. It would not be acceptable if it only included measures that are tightly aligned.

We suggest that these indicators be used to define “acceptable” and “high” quality research proposals and studies. To be considered acceptable quality, a research proposal or study would need to meet all but one of the Essential Quality

Indicators and demonstrate at least one of the quality indicators listed as Desirable as shown in Tables 1 and 2. To be considered high quality, a proposal or study would need to meet all but one of the Essential Quality Indicators and demon-

strate at least four of the quality indicators listed as Desirable. These definitions of acceptable and high quality are tentative and should be field-tested by universities, agencies that review grant applications, and research organizations.

In the following section, we walk the reader through the framework by providing brief descriptions of the indicators. Finally, we suggest a standard for determining whether a specific practice may be considered evidence based.

QUALITY INDICATORS FOR GROUP EXPERIMENTAL AND QUASI-EXPERIMENTAL RESEARCH

CONCEPTUALIZATION OF THE RESEARCH STUDY

Because a research study without context will have no impact, it is critical that the researcher clearly present the conceptualization and design of a study. In this section, we describe four indicators for providing context for a research study.

The conceptualization of the research study is based on the findings of rigorously designed studies that reflect the current scope of extant knowledge, including the findings of seminal studies. OR If an innovative approach is proposed, it is based on sound conceptualization and is rooted in sound research.

The literature review is a key section for the conceptual design of a study. Typically, it describes prior research that leads to current conceptualizations and study design. It is important that the review presents the existing information and makes the case for the proposed research. The review of literature should reflect both recent and seminal research studies in the area, making the reader aware of how these studies relate to the proposed research study. If there is no recent research, the researcher should state this clearly.

Researchers should ensure that the review of literature not only is adequate in the breadth of studies covered, but also is focused on the small set of issues that are critical for this particular study. The end result is that the researcher should present a concise but complete summary of the scientific knowledge base for the area in which

the problem exists, should illuminate areas of consensus and areas that require further investigation, and create a theoretical and conceptual understanding for why the study addresses an important topic that has not been fully addressed in the past. Whether the researcher proposes an innovative approach for which little existing empirical evidence exists, or interventions that are subtle variants on evidence-based practices, he or she should focus the review of literature on providing a compelling argument for the approach, and setting the stage for the guiding research questions.

A compelling case for the importance of the research is made.

Feuer and colleagues (2002) make a strong case for an increased emphasis on presenting arguments regarding the importance of any research study when writing or discussing research. Part of their argument focuses on the increased salience of educational research in the current political climate. The question of how researchers choose the topics on which they focus their research is important, and the rationale for these choices must be conveyed to consumers or funders of research projects.

Fabes, Matrin, Hanish, and Updegraff (2000) recently discussed the issue of evaluating the significance of developmental research for the 21st century. These authors identified four new types of "validity" which, with some adaptation, can be applied to educational research. These types of validity are *incidence validity*, *impact validity*, *sympathetic validity*, and *salience validity*.

Incidence validity refers to the degree to which a particular piece of research addresses a topic that significantly affects large numbers of people. *Impact validity* is the degree to which a research topic is perceived to have serious and enduring consequences. *Sympathetic validity* reflects the tendency to judge the significance of the research based on the degree to which it generates feelings of sympathy for individuals affected by the problem under study. *Salience validity* reflects the degree to which people, generally referring to the public, are aware of the problem or topic. Fabes and colleagues (2000) note that it is difficult for one study to incorporate all four types of validity, but that it is important for researchers to

be aware of these types of validity and to look at the pattern of validity addressed by the study in question. These seem helpful concepts to use in considering this quality indicator.

Because a research study without context will have no impact, it is critical that the researcher clearly present the conceptualization and design of a study.

Valid arguments supporting the proposed intervention as well as the nature of the comparison group are presented.

It is important to characterize the intervention and place it within a context that makes sense to the reader. The rationale for the proposed intervention should be sound and based, in part, on facts and concepts presented in the review of the literature. If the intervention is applied to new participants, settings, or context, there should be clear links based on argument and/or research for this new use.

In addition to describing the intervention, it is important to describe the nature of the experience or services to be received by the comparison group (Feuer et al., 2002; Gersten, Baker, & Lloyd, 2000). Potentially, a variety of reasons may exist for establishing comparison groups. The most common is describing an innovative approach to a traditional one. In this case, readers need to know the nature of experiences of participants in the more traditional comparison group. At times, researchers contrast theoretically opposed or divergent interventions. This, too, is legitimate, especially when the contrast is focused and of high interest. There are occasions when a researcher is broaching a new topic and simply wants to know if it leads to change. These would be the only instances where a no-treatment comparison group would be appropriate. Designs that are more elegant include multiple groups, some involving subtle contrasts (e.g., taught by an interventionist or teacher vs. a paraprofessional; identical intervention with or without progress monitoring).

The research questions are appropriate for the purpose of the study and are stated clearly.

In most research proposals, a succinct statement of the key questions addressed is critical. These questions should be linked in an integral fashion to the purpose statement. The criteria for the research questions are simple and serve to pull the conceptualization and design of the study together. Hypothesized results should be specified, but the authors of a proposal should be candid about issues for which they have no specific prediction and reviewers need to be aware that it is equally appropriate to frame clear research questions without a specified hypothesis.

DESCRIBING PARTICIPANTS

Sufficient information to determine/confirm whether the participants demonstrated the disability(ies) or difficulties addressed is presented.

A fundamental issue for generalizing study results is whether the participants actually experienced the disability(ies) or the difficulties the study was intended to address. To resolve this issue, researchers must move beyond school district-provided labels. Informing readers that students met the state criteria for identifying students as developmentally delayed provides little information on the type or extent of disability a student may have. Researchers need to provide a definition of the relevant disability(ies) or difficulties and then include assessment results documenting that the individuals included in the study met the requirements of the definition. For example, if the study addressed students with math difficulty, the researchers would (a) define math difficulty (e.g., performing below the 16th percentile on a test of mathematics computation and performing below the 16th percentile on a test of mathematics concepts) and (b) document that each participant included in the study met these criteria. We recommend that researchers attempt to link their definitions to those in the current literature, which can be done by either replicating the definition used in prior published research or by explaining the reason for extending or redefining the operational definition, and provide a new label for this alternative definition (to

increase specificity of terms in research).

Researchers should provide enough information about participants so that readers can identify the population of participants to which results may be generalized. This additional information may include, but is not limited to, comorbidity disability status (e.g., what percentage of students with reading disabilities also had math disability and/or attention deficit/hyperactivity disorder?); demographics (e.g., age, race, sex, subsidized lunch status; English language learner status, special education status); scores-related academic assessments (with effect sizes also desirable); and percentage of students receiving subsidized lunch for participating schools.

As part of describing participants, group difference on salient variables must also be presented. Given the issues previously discussed on participant selection, comparability of groups on key demographic variables must be examined, both to describe the participants as well as to use in later analyses. It is also the researchers' responsibility to document sample comparability at pretest on at least one outcome measure (or key predictor of outcome). Demonstrating such comparability eliminates the possibility that study effects accrued because of preexisting differences between the study groups on key performance or demographic variables.

Appropriate procedures are used to increase the probability that participants were comparable across conditions.

The optimal method for assigning participants to study conditions is through random assignment, although in some situations, this is impossible. It is then the researchers' responsibility to describe how participants were assigned to study conditions (convenience, similar classrooms, preschool programs in comparable districts, etc.). Researchers are urged, however, to attempt random assignment. In our experience, some type of random assignment can often be negotiated with districts or school or clinic personnel. The quality of the research design is invariably higher with random assignment of both student participants and intervention providers.

However, there will be situations where random assignment of students is not feasible. In

these cases, random assignment of teachers or interventionists can be an excellent alternative, if used appropriately. Random assignment of classrooms, or even schools to treatment conditions is also a viable option. These are all legitimate design strategies for randomized trials if appropriate statistical analyses are used and researchers are aware of issues that can affect statistical power. For example, statistical analyses need to account for the nesting of students in classrooms. Therefore, power analyses may need to be conducted at both the student and the classroom level. One advantage of this type of design is that it is often not necessary to assess each student in a class. A randomly selected subsample is often appropriate. Another advantage is that, in some cases, power is increased because the standard deviation between classes is often much lower than the standard deviation between all the students. We recommend that researchers work closely with a skilled statistician on these issues.

It is important to note that random assignment of participants to study conditions does not guarantee equivalent study groups on key characteristics of the participants. In fact, given the heterogeneity of many disability subgroups, it can be difficult to achieve initial comparability on key variables, with or without random assignment. For this reason, matching participants on a salient variable(s) and randomly assigning one member of each pair to a treatment condition, or a stratified assignment procedure to study conditions is often preferred.

Researchers need to provide a definition of the relevant disability(ies) and then include assessment results documenting that the individuals included in the study met the requirements of the definitions.

Differential attrition among intervention groups or severe overall attrition is documented.

Researchers should document overall attrition of participants and ensure that attrition rates between the intervention and comparison groups were not substantially different. The reason for

this is to document that study groups remain comparable at the conclusion of the study.

Sufficient information describing important characteristics of the intervention providers is supplied, and appropriate procedures to increase the probability that intervention providers were comparable across conditions are used.

Researchers should supply enough information about the intervention providers (i.e., teachers or other individuals responsible for implementation) so that readers understand the type of individual who may be capable of administering the intervention when it is used outside the context of that study (i.e., the intervention providers to which results may be generalized). Relevant characteristics may include age, sex, race, educational background, prior experience with related interventions, professional experience, and number of children with and without disabilities in the family (for parents). It may also be useful to present information on assessments involving knowledge of the topic to be taught, knowledge of appropriate pedagogical methods, efficacy measures, and/or attitudinal measures.

The researchers must describe how intervention providers were assigned to the various study conditions. Random assignment is the preferred method, but other approaches often are necessary because of the circumstances in which the study is conducted. For example, researchers should also consider situations where intervention providers are counterbalanced across conditions. This can be done by having each interventionist teach one group with one method and another group with another method or by switching interventionists across conditions at the midpoint. The burden is on the researchers to describe assignment methods and, when random assignment or counterbalancing is not employed, to provide a rationale for how those methods support the integrity of the study.

Comparability of interventionists across conditions should be documented. Regardless of whether random assignment was employed, researchers must quantify relevant background variables for the intervention providers associated with each of the study conditions. This is especially true if the interventionists are not counter-

balanced across conditions. With such information, readers can be assured that the effects of the intervention are not due to preexisting differences between the intervention providers (i.e., the favorable effects for an intervention are in fact attributable to the intervention).

IMPLEMENTATION OF THE INTERVENTION AND DESCRIPTION OF NATURE OF SERVICES IN COMPARISON CONDITIONS

The intervention is clearly described and specified.

Researchers should provide a precise description of the independent variable to allow for systematic replication as well as to facilitate coding in research syntheses such as meta-analysis (Gersten, Baker, & Lloyd, 2000). Unfortunately, the instructional labels that are often assigned to interventions are vague or misleading and may vary considerably from study to study. For example, in a meta-analysis on the impact of various instructional approaches on students with learning disabilities, Swanson and Hoskyn (1998) found that there was significant overlap in the way constructs with similar labels were operationalized. They concluded that classifying types of teaching in broad terms such as direct instruction and strategy instruction was problematic and that more fine-grained terminology and descriptions needed to be used.

Interventions should be clearly described on a number of salient dimensions, including conceptual underpinnings; detailed instructional procedures; teacher actions and language (e.g., modeling, corrective feedback); use of instructional materials (e.g., task difficulty, example selection); and student behaviors (e.g., what students are required to do and say). Precision in operationalizing the independent variable has become 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 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.

Fidelity of implementation is described and assessed in terms of surface (the expected intervention is implemented) and quality (how well the intervention is implemented) features.

Fidelity of implementation (also known as treatment fidelity or treatment integrity) refers to the extent to which an intervention is implemented as intended (Gresham, MacMillan, Beebe-Frankenberger, & Bocian, 2000). Information about treatment fidelity is essential in understanding the relationship between an intervention (i.e., independent variable) and outcome measures (i.e., dependent variables). The goal of experimental research in special education is to demonstrate that any changes in a dependent variable are the direct result of implementing a specified intervention. Without evidence about whether the intervention was actually implemented as planned, however, it is impossible to establish this relationship unequivocally.

This indicator is concerned with both *whether* treatment fidelity was measured and *how* it was measured. Although necessary, it is not sufficient to only note the number of days/sessions that intervention was conducted. At the least, researchers should observe the intervention using a checklist of treatment components and record whether the most central aspects of the intervention occurred. A fidelity score can then be derived by calculating the percentage of treatment components that were implemented. Observations should take place over the entire course of the study on a regular basis. The research team should include some measure of interobserver reliability. Key features to look for in assessing this criterion are: (a) inclusion of key features of the practice (e.g., ensuring students use all five steps in the strategy, ensuring students are asked to draw a visual representation of a mathematical problem, ensuring teachers provide the specified number of models of a cognitive strategy); (b) adequate time allocation per day or week for the intervention; and (c) coverage of specified amount of material in the curriculum or teacher guide (when appropriate). Although we label this *surface fidelity*, it is an essential part of any study and, thus, extremely important.

Some interventions, such as home visiting interventions, may not be as “visible,” but still require fidelity assessments. Videotapes of visits may need to be obtained and scored. Occasionally interventions will theorize a causal chain that requires assessment of intermediate links in the chain. For example, a home visiting intervention may theorize that the home visitors working with parents in a certain way will change the way parents work with their child, which in turn will influence the child’s behavior. In this type of situation, assessment of the causal link, the way the parents work with their child, is important to be sure that the intervention worked as theorized.

More sophisticated measures of fidelity require an observer to document not only the occurrence of prescribed steps in the process, but also that the interventionist employed the techniques with quality. For example, an observer can evaluate whether teachers model skills and strategies using clear language and interesting examples, or whether they scaffold or provide corrective feedback consistently. Investigating treatment fidelity at this level provides a deeper understanding of implementation issues and can lead to important insights about intervention components and teacher behaviors that are more directly related to desired outcomes. It also can be a lens into how well interventionists understand the principles behind the concept. This level always requires some level of inference.

Use of observational field notes or audiotapes of select sessions can be used to gain an understanding of quality. Often, selected use of transcripts can enrich a journal article or research report on the study by giving the reader a sense of how the intervention plays out with students and of actual curricula materials (e.g., Fuchs & Fuchs, 1998; Palincsar & Brown, 1984). Finally, it is important to assess treatment fidelity in a way that allows researchers to ascertain whether the level of implementation was constant or whether fidelity varied or fluctuated within and across interventionists. Researchers should carry out this type of analysis both for individual interventionists over time to determine if implementation was consistent across the duration of the intervention as well as between interventionists to determine if some teachers delivered the treatment with greater integrity than did others.

These data can serve numerous other purposes. If researchers find a reasonable amount of variability, as will often be the case in studies that use quality of implementation measures, they can correlate implementation scores with outcomes or use contrasted groups analysis to see if quality or intensity of implementation relates to outcome measures. They also can use the profiles to get a sense of typical implementation problems. These data can be used to develop or refine professional development activities.

The nature of services provided in comparison conditions are described and documented.

One of the least glamorous and most neglected aspects of research is describing and assessing the nature of the instruction in the comparison group (Gersten, Baker, et al., 2000). Yet, to understand what an obtained effect means, one must understand what happened in the comparison classrooms. This is why researchers should also describe, assess, and document implementation in comparison groups.

At a minimum, researchers should examine comparison groups to determine what instructional events are occurring, what texts are being used, and what professional development and support is provided to teachers. Other 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). In some studies, assessment of the comparison condition may be similar to what occurs for treatment fidelity. Every study will vary based on specific research issues.

OUTCOME MEASURES

Multiple measures are used to provide an appropriate balance between measures closely aligned with the intervention and measures of generalized performance.

Far too often, the weakest part of an intervention study is the quality of the measures used to evaluate the impact of the intervention. Intervention researchers often spend more time on aspects of the intervention related to instructional

procedures than on dependent measures. However, using tests of unknown validity invariably weakens the power of a study. 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.

One of the foremost challenges in crafting a study is selection of measures that are well aligned with the substance and the intervention and yet are sufficiently broad and robust to (a) avoid criticism for "teaching to the test" through the specific intervention, and (b) 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 of generalized performance. Multiple measures should be used in a study, as no measure is perfect, and no measure can assess all or even most of the important aspects of performance that an intervention might affect. If a study addresses more than one construct (e.g., fluency with arithmetic facts and problem-solving ability), it can be valuable to include multiple tools to measure each facet of performance.

Finally, in special education research that includes a wide range of different types of students, it is critical that measures are able to adequately assess performance of students with different disabilities. This may sometimes involve use of accommodations or use of alternate formats such as structured interviews on the content rather than written essays (see for example Bottge, Heinrichs, Mehta, & Ya-Hui, 2002; Ferretti, MacArthur, & Okolo, 2001; Gersten, Baker, Smith-Johnson, Peterson, & Dimino, in press).

Evidence of reliability and validity for the outcome measures is provided.

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; however, this is often neglected. 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 helps us to understand how well a cluster of items on a test “fit” together, and how well performance on one item predicts performance on another. These analyses help researchers locate items that do not fit well, or items with a weak item-to-total correlation. Revising these items or dropping them from the measure used increases the reliability and thus increases the power of a study.

The bottom line for reliability are coefficient alpha reliabilities of at least .6. This may seem low; however, measurement and design experts such as Nunnally and Bernstein (1994) and Shadish et al. (2002), indicate that this level of reliability is acceptable for a newly developed measure or a measure in a new field for two reasons. First, in experimental research, as opposed to individual assessment, we are concerned with the error associated with the mean score. The standard error of measurement of the mean is appreciably less for a sample mean than an individual test score. Second, internal consistency of .6 or higher indicates that a coherent measurement construct is being measured.

An ideal mix of outcome measures in a study would also include psychometrically sound measures with a long record of accomplishment, such as the Woodcock Johnson or Walker-McConnell, with Cronbach alpha reliabilities over .8 and newer measures with coefficient alpha reliability of .6 or so. Although we strongly encourage inclusion of data on test–retest reliability, an intervention study is typically considered acceptable without such data, at the current point in time.

At the same time, some indication of concurrent validity is also essential. For newly developed measures, using the data generated in the study would be acceptable, as would data from a study of an independent sample. Yet, to be highly acceptable, the researcher should independently conduct some type of concurrent validity. Concurrent validity becomes even more critical when using measures for groups other than those for which the test was designed (e.g., using a measure translated into Spanish for use with Spanish-speaking bilingual students). For studies to rank in the highly acceptable category, empirical data on predictive validity of measures used and any information on construct validity should be reported.

Outcomes for capturing the intervention's effect are measured at the appropriate times.

The goal of measurement activities in experimental research is to detect changes in a dependent variable, which are the result of implementing an independent variable. Depending on the nature of the independent variable and a study's research questions, there may be critical periods of time in which intervention effects can be captured through data collection activities.

Many times, the effects of an intervention are best detected immediately. For example, if researchers are interested in evaluating the initial effectiveness of a 4-week intervention designed to introduce a strategy for deriving word meanings from context, data should be collected within a few days of the end of the intervention. If data collection is delayed, initial intervention effects may fade or subsequent instruction or other uncontrolled events may contaminate results.

Sometimes, however, important intervention effects cannot be detected immediately. Information about the delayed or long-term effects of interventions is extremely important but often not addressed or measured in research studies. For example, researchers may be interested in determining whether students are still able to apply the word-deriving strategy 6 weeks after the intervention ended. In cases such as this, researchers should time data collection activities to align with research questions.

A good deal of the researcher's effort should be devoted to selection and development of dependent measures.

For some studies, it may be appropriate to collect data at only pre- and posttest. In many cases, however, researchers should consider collecting data at multiple points across the course of the study, including follow-up measures. Multiple data points allow researchers to apply more complex statistical procedures such as Hierarchical Linear Modeling or Growth Curve Analyses. These analyses allow for an examination of indi-

vidual student trajectories as well as overall group effects.

These data can provide researchers with a more nuanced, complex picture of student growth as well as information about immediate and long-term intervention effects. Thus, we would consider these indicators of exemplary research designs, provided the array of measures makes sense developmentally.

Data collectors and/or scorers are blind to study conditions and equally (un)familiar to examinees across study condition.

Data collection activities may inadvertently affect a study's findings. One way this can happen is if data collectors and/or scorers are aware of which students took part in the treatment and control conditions. Experimenter bias may result from this situation if researchers' expectations unintentionally influence assessment or scoring procedures. Study findings can also be affected if some participants are familiar with data collectors while other participants are not. In this case, participants may feel more comfortable with individuals that they know and therefore perform better during testing. Researchers should design and carry out data collection activities in a way to minimize these threats to the study's internal validity.

Having data collectors/scorers blind to study conditions and examinees is optimal for any study, although there will be times when it is impossible to implement. In all cases, however, data collectors should be trained and interrater/observer/tester reliabilities conducted.

Adequate interscorer agreement is documented.

Although researchers must select outcome measures that align with the purpose of the study and that possess acceptable levels of reliability and validity, the quality of the data from these measures is only as good as the corresponding data collection and scoring procedures. To ensure the interpretability and integrity of study data, researchers should also consider carefully how data is collected and scored.

Researchers should ensure that test administration and scoring is consistent and reliable across all data collectors and scorers. Documenta-

tion of agreement is especially important when using newly developed measures or when scoring involves subjective decisions. Interscorer agreement can be evaluated by having multiple data collectors administer assessments to a random sample of study participants or by observing selected testing sessions. Similar procedures can be used to check agreement in data scoring. Ideally, reliability in data collection and scoring should be above .90.

DATA ANALYSIS

We highlight several key themes in this set of indicators. The linkage of data analysis and unit of analysis to key research questions is important. In addition, we emphasize that the researcher use appropriate statistical techniques to adjust for any pretest differences on salient predictor variables.

A key issue is ensuring that data analyses and research questions are aligned with the appropriate unit of analysis for a given research question. Researchers should actually define which unit was used in the statistical analysis of intervention effects. For example, in one type of analysis, individual students may be the unit of analysis; in another case, where students are taught in pairs, the pair may well be the appropriate unit of analysis.

The determination of the appropriate unit of analysis relates directly to the research hypotheses/questions, the sample, and assignment of sample to experimental conditions. When students are the unit, the n equals the number of students in the analysis and each student contributes his/her score. When classrooms are the units, the n equals the number of classrooms, and individual students' scores are averaged to reflect a class value. When a full class is a unit, it may well be advisable to assess all of the special education students but assess only a stratified random sample of the nondisabled students.

It is also possible that within the same study, for example, the teacher may be the appropriate unit of analysis for some measures (e.g., teacher knowledge of inclusive practices), and the individual student may be the best unit of analysis for others (e.g., degree of engagement in academic activities with peers). When appropriate, we urge the use of multilevel analyses because

they are uniquely designed to consider multiple units within a single analysis.

The quality indicators related to data-analytic strategies and study design are essential. However, it would require, at a minimum, a full issue of this journal to begin to do justice to the topic. Choice of appropriate data-analytic strategy has always been partly art and partly science. Usually, numerous strategies are appropriate, some more elegant and sophisticated than others. In the past decade, there have been enormous advances in this area, with increased understanding of and wide use of structural equation modeling and multilevel modeling. Thus, the array of options is large and determining a strategy or set of strategies that is sensitive/powerful enough to detect treatment effects, while also sophisticated enough to look at important secondary patterns or themes in the data, is a complex task.

In this brief section, we highlight a few of the key themes that both those designing studies and those evaluating proposals or research reports need to keep in mind when considering data analyses.

The data analysis techniques chosen are appropriate and linked in an integral fashion to key research questions and hypotheses.

Statistical analyses are accompanied with presentation of effect sizes. Merely cataloguing every possible analysis that can be done is unacceptable. A brief rationale for major analyses and for selected secondary analyses is critical.

The variability within each sample is accounted for either by sampling or statistical techniques such as analysis of covariance.

Postintervention procedures are used to adjust for differences among groups of more than 0.25 of a standard deviation on salient pretest measures. Planned comparisons are used when appropriate.

The researcher should clearly link the unit of analysis chosen to the key statistical analyses.

Support for the unit(s) of analysis relative to the conceptual and practical foci of the treatment is made. Multilevel analysis is used when appropriate. Although data can always be ana-

lyzed at multiple levels, the reader of a proposal or report should clearly know the levels for which the analysis will be most powerful and which unit of analysis is most appropriate for a specific research question and design. Appropriate secondary analyses are conducted (e.g., testing for the effectiveness of the treatment within important subgroups of participants or settings).

A power analysis is provided to describe the adequacy of the minimum cell size. A power analysis is conducted for each unit of analysis to be examined (e.g., school and class as well as student).

If an intervention is effective for the population targeted, statistical power denotes the probability that a statistical test will accurately denote statistical significance. Factors traditionally identified as determining statistical power include, sample size, anticipated effect size, and type of statistical test selected. Given the increasing use of multilevel models, we would like to add the number of assessment waves as an additional factor affecting statistical power.

Under conditions of study economy (as is often the case in special education intervention research), the appropriate number of units needed in the design is determined by the effect size one expects to detect (small vs. medium vs. large). Cohen's (1988) traditional break out of effect size ranges is: .2 = small, .5 = moderate, and .8 and greater = large. Effect sizes in the .40 or larger range are often considered minimum levels for *educational* or *clinical* significance (Forness, Kavale, Blum, & Lloyd, 1997; Walberg, 1986). Experienced special education investigators determine the number of participant units needed in a study using effect sizes determined from pilot data that have used the intervention and relevant outcome variables. Alternatively, one may obtain expected effect sizes from related published research for the purpose of conducting a power analysis and planning a new 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. In early stages of research on a topic, it makes sense to design a study conservatively, with a large enough sample to increase the chance of detecting small effects. Conducting a study in two or even

three waves or cycles often makes sense in these instances.

In some cases, the costs of designing a study capable of detecting small effects may simply not be interesting or worth it. Thus, smaller participant units can be used. The point is that special education researchers can conduct studies that are correctly powered relative to available resources, student population sizes, and goals of the research when hypothesizing moderate-to-large effects and keeping the number of experimental conditions small (i.e., two rather than three group design). Replicating a series of studies with smaller samples powered to detect moderate-to-large effects appears to offer a powerful approach to the discovery of causal relations in small populations/samples.

When proposing a study, researchers should provide a power analysis for the most important analyses and either explicitly or implicitly indicate how they address the major factors that influence the power of a research design. In other words, they need to indicate whether they anticipate a small, medium, or large effect and how they have addressed issues of sample heterogeneity/variability in their design and analysis strategies. The power analysis should indicate that the design has adequate sample size to detect hypothesized effects. The researchers should indicate the impact of waves of assessments on the power of their analyses.

DETERMINING WHEN A PRACTICE IS EVIDENCE-BASED: A MODEST PROPOSAL

Currently, there is a great deal of deliberation and discussion on what it means to call an educational practice or special education intervention *evidence based*. Particularly controversial issues are the relative weightings of randomized trials versus quasi-experiments and to what extent we can generalize the findings across various subgroups of students with disabilities. Another key issue is how to make the determination that an evidence-based practice is implemented with such low quality that we could no longer assert that it is likely to enhance learner outcomes.

The authors of this article were not in complete agreement on any of these issues. However,

we decided, with some trepidation, to present a modest proposal, a criterion for concluding a practice is *evidence based*. In reaching this determination, we were heavily influenced by the research and scholarship on research synthesis and meta-analysis over the past 20 years (e.g., Cooper & Hedges, 1994), and the research syntheses conducted in special education (e.g., Gersten, Schiller, & Vaughn, 2000; Swanson & Hoskyn, 1998).

- There are at least four acceptable quality studies, or two high quality studies that support the practice; *and*
- The weighted effect size is significantly greater than zero.

Again, to be considered acceptable, a study would need to meet all but one of the Essential Quality Indicators specified in Table 2 and demonstrate at least one of the quality indicators listed as Desirable. To be considered high quality, a study would need to meet all but one of the Essential Quality Indicators specified in Table 2 and demonstrate at least four of the quality indicators listed as Desirable.

The reason we stressed the weighted effect size is that this statistic takes into account (a) the number of studies conducted on the specific intervention, (b) the number of participants in the studies, and (c) the magnitude and consistency of effects. To us, these three components seem to be most critical for making determinations as to whether a practice is evidence based.

We propose the following criteria for considering a practice as *promising*:

- There are at least four acceptable quality studies, or two high quality studies that support the practice; *and*
- There is a 20% confidence interval for the weighted effect size that is greater than zero.

To reach these recommendations, we weighed the quality of the research design (e.g., Feuer et al., 2002; Shadish et al., 2002) heavily. There are several reasons for doing this. Recently, Simmerman and Swanson (2001) documented the impact of specific flaws in research design on a study's effect size. They argued, "Among the factors that determine a study's credibility, it is

often the internal and external validity variables that establish research quality and have significant effects on treatment outcomes” (p. 211). To test this hypothesis, Simmerman and Swanson examined the effects of a large set of internal and external validity variables on treatment outcomes for students with learning disabilities (LD) using studies identified in the Swanson and Hoskyn (1998) meta-analysis on treatment outcomes and LD.

Results indicate the following factors lead to lower effect sizes: (a) controlling for teacher effects, (b) using standardized rather than just experimenter-developed measures, (c) using the appropriate unit in data analyses, (d) reporting the sample’s ethnic composition, (e) providing psychometric information, and (f) using multiple criteria defining the sample. Simmerman and Swanson’s (2001) results indicate that better controlled studies appear to be less biased in favor of the intervention. It is interesting to note that use of random assignment (versus use of quasi-experimental designs) approached, but did not reach, statistical significance ($p = .0698$).

Thus, issues addressed by these quality indicators appear to influence the inferences we make about effective approaches for special education. Better controlled studies (in terms of using measures of documented reliability, controlling for effects of teachers or interventionists, and using the appropriate unit of analysis in statistical calculations) tend to be less biased. Use of standardized measures of broad performance lead to weaker estimates of effectiveness of an intervention. The key lesson we can learn from the Simmerman and Swanson (2001) analysis is that research quality does matter and does have educational implications.

We feel it is important to note that conducting quality research is more expensive than conducting experiments that are compromised at various levels. We see signs of awareness of this fact in both the National Research Council (2002) report on educational research and in some current federal initiatives. Routinely conducting quality research in public schools will also require a shift in the culture of schools, much as it required a shift in the culture of medical clinics and hospitals 50 years ago, and of public welfare programs 2 decades ago. Active support by the U.S. Depart-

ment of Education will, in our view, be required. It would seem to go hand-in-hand with the current emphasis on scientifically based research.

This set of criteria is merely a first step. We envision the following as critical next steps:

We feel it is important to note that conducting quality research is more expensive than conducting experiments that are compromised at various levels.

- Field-testing of this system of indicators by competent individuals as they review grant applications and other research proposals and manuscripts submitted to journals for publication.
- Refinements based on field-testing in both the areas of proposal review and of research synthesis.
- Consideration for adoption by journals in our field and/or funding agencies such as Institute of Educational Sciences and Office of Special Education Programs.
- Serious field-testing of the quality indicators’ impact on evidence-based practice needs to be conducted. The issue of integrating findings from different types of research (e.g., correlational, single-subject design, qualitative) needs to be considered as part of this field-testing effort.

This would seem to be a reasonable effort for CEC’s Division for Research to undertake in the near future in conjunction with relevant federal agencies, other divisions of CEC, and, potentially, agencies that review research grant applications relating to special education.

REFERENCES

- Boruch, R. F. (1997). *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: Sage.
- Bottge, B. A., Heinrichs, M., Mehta, Z. D., & Ya-Hui, H. (2002). Weighing the benefits of anchored math in-

- struction for students with disabilities in general education classes. *Journal of Special Education*, 35, 186–200.
- Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cooper, H., & Hedges, L. V. (Eds.). (1994). *Handbook of research synthesis*. New York: Russell Sage Foundation.
- Elbaum, B., Vaughn, S., Hughes, M., & Moody, S. W. (1999). Grouping practices and reading outcomes for students with disabilities. *Exceptional Children*, 65, 399–415.
- Fabes, R. A., Matrin, C. L., Hanish, L. D., & Updegraff, K. A. (2000). Criteria for evaluating the significance of developmental research in the twenty-first century: Force and counterforce. *Child Development*, 71, 212–221.
- Ferretti, R. P., MacArthur, C. D., & Okolo, C. M. (2001). Teaching for historical understanding in inclusive classrooms. *Learning Disability Quarterly*, 24, 59–71.
- Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational Researcher*, 31, 4–14.
- Forness, S., Kavale, K. A., Blum, I. M., & Lloyd, J. W. (1997). Mega-analysis of meta-analyses: What works in special education and related services. *TEACHING Exceptional Children*, 29, 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.
- Gall, M. D., Borg, W. R., & Gall, J. P. (2002). *Educational research: An introduction* (7th ed.). White Plains, NY: Pearson/Allyn & Bacon.
- Gersten, R., Baker, S., & Lloyd, J. W. (2000). Designing high quality research in special education: Group experimental design. *Journal of Special Education*, 34, 2–18.
- Gersten, R., Baker, S., Smith-Johnson, J., Peterson, A., & Dimino, J. (in press). Eyes on the prize: Teaching history to students with learning disabilities in inclusive settings. *Exceptional Children*.
- Gersten, R., Schiller, E. P., & Vaughn, S. (Eds.). (2000). *Contemporary special education research: Syntheses of the knowledge base on critical instructional issues*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Gresham, F. M., MacMillan, D. L., Beebe-Frankenberg, M. E., & Bocian, K. M. (2000). Treatment integrity in learning disabilities intervention research: Do we really know how treatments are implemented? *Learning Disabilities Research & Practice*, 15, 198–205.
- National Research Council. (2002). *Scientific research in education*. (R. J. Shavelson & L. Towne, Eds.), Committee on Scientific Principles for Educational Research. Washington, DC: National Academy Press.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, 1, 117–175.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for general causal inference*. Boston: Houghton Mifflin.
- Simmerman, S., & Swanson, H. L. (2001). Treatment outcomes for students with learning disabilities: How important are internal and external validity? *Journal of Learning Disabilities*, 34, 221–236.
- 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.
- Valentine, J. C., & Cooper, H. (2003). What Works Clearinghouse Study Design and Implementation Assessment Device (Version 1.0). Washington, DC: U.S. Department of Education.
- Walberg, H. J. (1986). Synthesis of research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 214–229). New York: Macmillan.

ABOUT THE AUTHORS

RUSSELL GERSTEN (CEC #108), Director, Instructional Research Group, Signal Hill, California. **LYNN S. FUCHS** (CEC #185), Professor, and **DONALD COMPTON** (CEC TN Federation), Assistant Professor, Vanderbilt University, Nashville, Tennessee. **MICHAEL COYNE** (CEC CT Federation), Assistant Professor, University of Connecticut, Storrs. **CHARLES GREENWOOD** (CEC #436), Director, Juniper Gardens Childrens Project, University of Kansas, Lawrence. **MARK S. INNOCENTI** (CEC #512), Research Associate Professor, Utah State University, Logan.

Address all correspondence to Russell Gersten, Instructional Research Group, 2525 Cherry Ave., Suite #300, Signal Hill, CA 90755. E-mail: rgersten@inresg.org

We wish to acknowledge the excellent editorial feedback provided by Susan Marks, Madhavi Jayanthi, Jonathan R. Flojo, and Michelle Spearman.

Manuscript received August 2003; accepted May 2004.