Determining Evidence-Based Practices in Special Education

BRYAN G. COOK
University of Hawaii at Manoa

MELODY TANKERSLEY
Kent State University

TIMOTHY J. LANDRUM
University of Virginia

ABSTRACT: Determining evidence-based practices is a complicated enterprise that requires analyzing the methodological quality and magnitude of the available research supporting specific practices. This article reviews criteria and procedures for identifying what works in the fields of clinical psychology, school psychology, and general education; and it compares these systems with proposed guidelines for determining evidence-based practices in special education. The authors then summarize and analyze the approaches and findings of the 5 reviews presented in this issue. In these reviews, prominent special education scholars applied the proposed quality indicators for high-quality research and standards for evidence-based practice to bodies of empirical literature. The article concludes by synthesizing these scholars' preliminary recommendations for refining the proposed quality indicators and standards for evidence-based practices in special education, as well as the process for applying them.

Such forces as the standards-based education movement, the mandated participation of students with disabilities in state proficiency testing, inclusion, and the recognition that many students with disabilities are capable of higher levels of academic and social attainment than previously expected have driven an intensified focus on improving outcomes for students with disabilities. Perhaps because many factors that may inhibit the outcomes of students with disabilities are beyond the direct control of educators (e.g., poverty, limited resources, attitudes), special educators have tended to focus their attention on one determinant of students' outcomes over which they have always exercised primary control—teaching practices. Unfortunately, many teachers of students with disabilities have implemented teaching practices shown to have little effect on student outcomes while eschewing many research-based practices (e.g., B. G. Cook & Schirmer, 2003; Kauffman, 1996). In an effort to bridge this research-to-practice gap, lawmakers have emphasized practices that research has shown to be effective in such legislation as the No Child Left Behind Act of 2001 and the Individuals With Disabilities Education Act of 2004. Researchers
in special education have also conducted initial research on how to effectively support teachers in adopting and maintaining the use of research-based or evidence-based practices (see Wanzek & Vaughn, 2006, for a review of this research). Despite the considerable interest in basing instructional practices on research evidence, special educators have not yet established definitively which practices are or are not evidence-based or settled on a systematic process for determining evidence-based practices (EBPs).

**Evidence-Based Practices**

All interventions are not equal; some are much more likely than others to positively affect student outcomes (Forness, Kavale, Blum, & Lloyd, 1997). Simple logic appears to suggest that, in general, teachers should prioritize the use of instructional practices that are most likely to bring about desired student outcomes. Although some contend that research cannot reliably determine which educational practices produce desired gains in student outcomes (e.g., Gallagher, 1998), we proceed under the positivist assumption that it can (Lloyd, Pullen, Tankersley, & Lloyd, 2006). The use of EBPs, or those practices shown by research to work, seems particularly imperative in special education. As Dammann and Vaughn (2001) suggested, whereas many nondisabled students make adequate progress under a variety of instructional conditions, students with disabilities require the most effective teaching techniques to succeed. However, advocating for implementing EBPs in special education begs two critical questions: what are EBPs, and how can researchers identify them?

**All interventions are not equal; some are much more likely than others to positively affect student outcomes.**

Determining whether a practice is evidence-based involves a number of issues: What types of research designs should researchers consider? How many studies with converging findings are necessary to instill confidence that a practice is effective? How methodologically rigorous must a study be for the results to be meaningful? To what extent must an intervention affect student outcomes for researchers to consider it effective? Although other issues certainly affect the difficult business of determining EBPs, we limit our discussion to these four issues—research design, quantity of research, methodological quality, and magnitude of effect.

**Research Design**

A generally accepted tenet of educational research holds that research designs exhibiting experimental control most appropriately address the question of whether a practice works (B. G. Cook, Tankersley, Cook, & Landrum, 2008). We recognize that no research design can completely rule out all alternative explanations for findings when conducted in the real-world settings of schools and classrooms; however, some designs do so more meaningfully than others. By using a control group, randomly assigning participants to groups, and actively introducing the intervention to the experimental group, group experimental designs can produce reliable knowledge claims regarding whether an intervention affects student outcomes (L. H. Cook, Cook, Landrum, & Tankersley, 2008). We are not implying that experimental research is better than other research designs; rather, different types of research address different questions, and researchers should use them accordingly. Should true experiments be the only research design considered in determining EBPs? Can quasi-experiments, single-subject research (SSR), correlational research, and qualitative research also meaningfully determine whether a practice works?

**Quantity of Research**

The process of conducting educational research and accumulating knowledge from research is tentative and cumulative (Rumrill & Cook, 2001). Because of the recognized vagaries in conducting field-based educational research (Berliner, 2002), it seems unwise to place too much faith in the results of a single study regardless of its design, effect size, or methodological rigor. Certainly, as more studies with converging evidence accrue, research consumers can have greater confidence in those findings. But how many studies
supporting a practice are sufficient to reasonably conclude that it works?

**Methodological Quality**

The methodological rigor with which a study is conducted affects the confidence that one can have in its findings. For example, evidence of acceptable implementation fidelity seems to be a necessary feature of a trustworthy study. If the researchers did not implement the intervention as designed, they can draw no meaningful conclusion about the effectiveness of the practice. Indeed, Simmerman and Swanson (2001) reported that the presence of desirable methodological features in a study (e.g., controlling for teacher effects, using appropriate units of analysis in analyzing data, reporting psychometric properties of measurement tools) significantly corresponds with lower effect sizes. Examining and accounting for the methodological quality of studies in determining EBPs therefore appears important. Should researchers determine EBPs by using only studies of high methodological quality? What methodological features are critically important for a high-quality study?

**Magnitude of Effect**

EBPs should have a considerable and meaningful—as opposed to trivial—positive effect on student outcomes. Researchers have traditionally gauged the impact of an intervention in group studies by using tests of statistical significance, which estimate the likelihood that differences between the groups occurred by chance. However, in part because of concerns that studies involving a large number of participants can yield statistically significant findings even when outcomes may not be educationally meaningful, researchers have begun to report effect sizes (e.g., Cohen’s *d*), which sample size does not affect, to help interpret an intervention’s effect (American Psychological Association, 2001). Although Cohen (1988) suggested values for interpreting effect sizes as small, medium, and large, he was careful to point out that researchers should not consider these subjective guidelines to be absolute standards. How is the effect of an intervention best evaluated? If researchers use effect sizes to assess the impact of a practice, how large an effect is necessary to indicate a meaningful change? If researchers use SSR studies to determine EBPs in special education, how should they evaluate the effect of the intervention?

The importance of using practices shown by research to be the most effective is by no means unique to special education (Odom et al., 2005). The medical field generally receives credit for pioneering efforts in this area, with evidence-based medicine becoming prominent in the 1990s (see Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996). Such other professions as clinical psychology (see Chambliss et al., 1996, 1998); school psychology (see Kratochwill & Stoiber, 2002; Task Force on Evidence-Based Interventions in School Psychology, 2003); and general education (see What Works Clearinghouse, WWC, n.d.a) have followed suit, developing criteria and procedures for identifying EBPs in their fields. To contextualize efforts to determine EBPs in special education, we briefly review the criteria and standards for determining EBPs in these three fields.

**Determining Evidence-Based Practices in Related Fields**

The Division 12 (Division of Clinical Psychology) Task Force on Promotion and Dissemination of Psychological Procedures (1995) delineated criteria, which Chambliss et al. updated in 1996 and 1998, for well-established treatments and probably efficacious treatments in clinical psychology. Subsequently, in the field of school psychology, Division 16 and the Society for the Study of School Psychology Task Force developed a detailed system for coding and describing multiple aspects of research studies (Task Force on Evidence-Based Interventions in School Psychology, 2003). Instead of categorizing the degree to which interventions are evidence-based, the coding system generated by the school psychology team provides a detailed description of a research base, from which consumers “draw their own conclusions based on the evidence provided” regarding the sufficiency of research supporting an intervention (Kratochwill & Stoiber, 2002, p. 360).

In general education, the WWC, established in 2002 by the U.S. Department of Education’s Institute of Education Sciences, rates reviewed
practices as having positive, potentially positive, mixed, no discernible, potentially negative, or negative effects (WWC, n.d.b). This section examines how these three diverse approaches for identifying what works in fields closely related to special education treat the issues of research design, quantity of research, methodological quality, and magnitude of effect.

**Research Design**

**Clinical Psychology.** Chambless et al. (1998) considered only studies employing between-group experimental and SSR designs in determining both well-established treatments and probably efficacious treatments.

**School Psychology.** The Task Force on Evidence-Based Interventions in School Psychology (2003) aims to provide descriptions of group research, SSR, confirmatory program evaluation, and qualitative research (Kratochwill & Stoiber, 2002). Coding manuals are currently available for group research and SSR, but are being expanded to include criteria for qualitative research and confirmatory program evaluation (T. Kratochwill, personal communication, September 26, 2008). Kratochwill and Stoiber suggested that coding nonexperimental research studies (i.e., qualitative and confirmatory program evaluation) provides information on a broad range of research relevant to consumers but do not indicate that these different research designs contribute equally to determining whether a practice works.

**General Education.** The WWC (n.d.b) requires at least one or two studies for a practice or curriculum to be considered as having positive, potentially positive, mixed, potentially negative, or negative effects. The specific number and type of studies required varies within and between these categories of effectiveness. For example, a positive effect requires two or more studies showing statistically significant positive effects, at least one of which meets WWC evidence standards without reservations, and no studies showing statistically significant or substantively important negative effects. A potentially positive effect, however, requires at least one study showing a statistically significant or substantively important positive effect, no studies showing statistically significant or substantively important negative effects, and no more studies showing indeterminate effects than studies showing statistically significant or substantively important positive effects.

**Methodological Quality**

**Clinical Psychology.** In addition to stipulating that researchers must compare interventions with a placebo or other treatment, the Division 12 criteria for well-established treatments require that
researchers (a) conduct experiments with treatment manuals, (b) clearly describe participant characteristics, and (c) have two separate investigators or investigatory teams conduct supporting studies (Chambless et al., 1998). These standards are relaxed for possibly efficacious treatments. Group experiments that compare the treatment group with a wait-list control group and that the same investigators conduct may be considered for possibly efficacious practices, as can SSR studies that the same investigators conduct.

When evidence regarding the effects of an intervention is mixed, reviewers further assess the methodological quality of studies to determine which studies to weigh more heavily (Chambless et al., 1998). Chambless and Hollon (1998) recommend assessing such methodological features as the following:

- The descriptions of samples use standard diagnostic labels assigned from a structured diagnostic interview.
- Outcome measures demonstrate acceptable reliability and validity in previous research.
- With the exception of simple procedures, the researchers follow a written treatment manual when delivering the intervention.
- Researchers avoid Type I error (e.g., adjust alpha level when conducting multiple statistical tests), control for pretest scores when comparing groups' posttest measures, and adjust analysis and interpretation if differential attrition or participation rates exist between groups.
- A stable baseline, typically with at least three data points, is established in SSR.

School Psychology. Although the Division 16 procedures do not classify studies according to their methodological quality, reviewers do rate and describe a number of methodological features—which consumers use to make informed decisions about an intervention's evidence base and effectiveness (Kratochwill & Stoiber, 2002). Reviewers evaluate studies, regardless of design, by using multiple criteria along three dimensions: general characteristics, key evidence components, and other descriptive or supplemental features. For example, researchers rate the strength of eight key components for group research on a 4-point scale. These key components are measurement, comparison group, statistical significance of outcomes, educational and clinical significance, implementation fidelity, replication, site of implementation, and follow-up assessment. In addition to providing an overall rating for each component, reviewers record additional information for most components. Regarding the comparison group, for example, reviewers select the type of comparison group from a list of options; rate their confidence in determining the type of comparison group (from very low to very high); indicate how the researchers counterbalanced change agents (by change agent, statistical, other); check how the researchers established group equivalence (e.g., random assignment, post hoc matched set, statistical matching, post hoc test for group equivalence); and check whether and how mortality was equivalent between groups.

General Education. The WWC (2008) specifies that for randomized controlled trials to meet evidence standards without reservations, (a) researchers must randomly assign participants to conditions; (b) overall and differential attrition must not be high; (c) no evidence of intervention contamination (e.g., changed expectancy, novelty, disruption, local history event) exists; and (d) researchers avoid a teacher-intervention confound by either assigning more than one teacher to each condition or by presenting evidence that teacher effects are negligible. The WWC uses similar, but less stringent, criteria for randomized controlled trials and quasi-experimental studies to meet evidence standards with reservations.

Magnitude of Effect

Clinical Psychology. For a group design study to support a well-established or possibly efficacious treatment, Chambless et al. (1998) require that treatment groups achieve outcomes that are statistically significantly superior to a control group or equivalent to a comparison group that received a treatment that researchers had previously determined to be well-established. With regard to SSR, Chambless and Hollon (1998) suggest that "evaluators . . . carefully examine data graphs and draw their own conclusions about the efficacy of the intervention" (p. 13).
School Psychology. Because the Division 16 Task Force (Task Force on Evidence-Based Interventions in School Psychology, 2003) coding procedures do not classify interventions in terms of their effectiveness, no criteria are specified regarding magnitude of effect. However, reviewers code study characteristics related to significance of outcomes: statistical significance, educational and clinical significance, and effect size for group studies; and visual analysis, effect size, and educational and clinical significance for SSR.

General Education. The WWC (n.d.b) uses five categories to describe the magnitude of effect for reviewed studies: statistically significant positive effects, substantively important positive effects, indeterminate effects, substantively important negative effects, and statistically significant negative effects. Substantively important effects are educationally meaningful although not statistically significant; the WWC suggests using an effect size of greater than ±0.25 as a cutoff for substantively important effects. Indeterminate effects are neither statistically significant nor have effect sizes greater than ±0.25.

Critiques of Processes for Determining What Works in Other Fields

Although it is difficult to disagree with the general notion that “evidence should play a role in educational practice” (Slavin, 2008, p. 47), controversy seems to follow closely on the heels of proposals for establishing EBPs. Indeed, Kendall (1998) likened EBPs to religion and politics as lightning rods for conflict. Elliott (1998) noted that criticisms of EBPs tend to fall into one of two categories: concerns about the general endeavor of designating EBPs and disagreements with the particular standards and criteria used. Although the first category includes many important issues (e.g., Can research conclusively identify any practice as truly effective? Will approaches not labeled as evidence-based be disregarded?), this article focuses here on critiques of specific features of the three processes reviewed.

Waehler, Kalodner, Wampold, and Lichtenberg (2000) noted that some have criticized the Division 12 criteria for determining empirically validated treatments in clinical psychology for relying too heavily on randomized clinical trials, psychological diagnoses, and adherence to treatment manuals, as well as for being too lenient. Scholars in school psychology also took issue with the Division 16 coding procedures as overwhelming and overly complex (Durlak, 2002; Levin, 2002; Nelson & Epstein, 2002; Stoiber, 2002); as seeming to endorse research designs that do not permit making causal inferences (Nelson & Epstein); and for producing ambiguous, descriptive reports rather than designating EBPs (Wampold, 2002). Finally, some researchers have criticized the WWC’s (2008) standards as relying too heavily on randomized controlled trials, which are extremely difficult to conduct in school settings (Kingsbury, 2006); as overly rigorous, resulting in few practices with positive effects identified (causing some to refer to the WWC as the “nothing works’ clearinghouse,” Viadero & Huff, 2006, p. 8); and as politically influenced (Schoenfeld, 2006).

Criticism regarding criteria and standards for determining EBPs may be unavoidable. Establishing EBPs involves addressing a number of questions that lack any unequivocally correct answers and about which different stakeholders are bound to disagree. For example, requiring a large number of randomized controlled trials that meet stringent methodological criteria and report large effect sizes will produce a high degree of confidence in practices shown to be evidence-based. However, this approach may be unnecessarily stringent, potentially excluding meaningful studies. Yet designating practices as evidence-based because of one study or a few research studies of any design without stringent methodological standards invites false positives.

The categorization of practices represents another contentious issue for which multiple valid approaches may exist. Using a dichotomous system for labeling practices (e.g., evidence-based or not evidence-based) provides straightforward input for prioritizing instructional practices. However, a binary categorization scheme may overlook the complexities involved in interpreting bodies of research literature as well as promote the unfounded view that practices are either completely effective or completely ineffective. In contrast, whereas in-depth descriptions of a research base might facilitate nuanced and comprehensive
understanding, they may be of limited practical use for practitioners seeking guidance on how to teach in their classrooms the following day.

Any approach to determining what works in special education will inevitably have limitations. This recognition does not suggest that endeavors to establish EBPs are destined to fail. Rather, the strength of a system for determining EBPs lies in matching criteria and standards with the collective traditions, values, and goals of the field that will use it. Therefore, special educators should design a system for determining what works in special education based on the unique characteristics and needs of their field. Odom et al. (2005) endeavored to delineate the “devilish details” (p. 138) of guidelines for determining EBPs rooted in the history and research traditions of special education.

PROPOSED GUIDELINES FOR EVIDENCE-BASED PRACTICES IN SPECIAL EDUCATION

As an initial step for basing practice on research, the Division for Research of the Council for Exceptional Children, under the leadership of Sam Odom, commissioned a series of papers that proposed quality indicators (QIs; i.e., features present in high-quality research studies) for four different research designs: group experimental studies (Gersten et al., 2005); SSR (Horner et al., 2005); correlational research (Thompson, Diamond, McWilliam, Snyder, & Snyder, 2005); and qualitative research (Brantlinger, Jimenez, Klingner, Pugach, & Richardson, 2005). Gersten et al. also proposed standards for determining EBPs on the basis of group experimental/quasi-experimental research, and Horner et al. proposed standards for determining EBPs on the basis of SSR. Considered together, the proposed QIs and standards constitute initial guidelines for establishing EBPs in special education. The number of prominent special education researchers who developed the proposed criteria and standards and the incorporation of feedback from special education researchers who discussed the proposed criteria and standards at a Research Project Director’s Meeting (hosted by the Office of Special Education Programs; Odom et al., 2004) enhances their credibility.

The following sections examine the proposed guidelines for determining EBPs in special education and compare the proposed guidelines in special education with the systems for determining what works in clinical psychology, school psychology, and general education in relation to research design, quantity of research, methodological quality, and magnitude of effect.

RESEARCH DESIGN

We assume that because standards for EBPs were proposed only for group experimental and quasi-experimental research (Gersten et al., 2005) and SSR (Horner et al., 2005), these research designs are the only ones to consider in determining whether a practice in special education is evidence-based. The Division for Research Task Force probably based this decision on the unique ability of these designs to exhibit experimental control (Cook, Tankersley, Cook, & Landrum, 2008). Special education, clinical psychology, and general education share many similarities in their treatment of research design in determining EBPs. For example, all three fields consider group experimental studies in determining EBPs. Researchers can also consider practices as evidence-based in special education, as well-established in clinical psychology, and as having potentially positive effects (but not as having positive effects) in general education on the basis of SSR. However, whereas Gersten et al. allowed for quasi-experimental studies to constitute the sole research support for EBPs in special education, Chambless et al. (1998) did not consider quasi-experimental research in establishing empirically validated therapies in clinical psychology, and the WWG (n.d.b) requires at least one true experiment to support practices with positive effects in general education.

QUANTITY OF RESEARCH

Gersten et al. (2005) required a minimum of two high-quality group studies or four acceptable-quality group studies to consider a practice evidence-based or promising in special education. These numbers are similar to the quantity of group-design studies required for determining
EBPs in clinical psychology and general education. For example, Chambless et al. (1998) required two or more group studies for a well-established treatment, and the WWC (n.d.b) calls for two or more group design studies, at least one of which must be a randomized controlled trial, to support practices with positive effects.

To consider a practice to be evidence-based in special education, Horner et al. (2005) specified a minimum of five SSR studies that involve a total of at least 20 total participants and that at least three different researchers conduct across at least three different geographical locations. This number is somewhat less than the number of SSR studies \((n = 9)\) that Chambless et al. (1998) required to deem a treatment in clinical psychology well established. By contrast, the WWC (2008) considers SSR studies as quasi-experimental designs, which cannot alone constitute sufficient evidence to deem a practice as having positive effects.

**Methodological Quality**

Gersten et al. (2005) proposed four essential QIs for group experimental research in the areas of describing participants, implementing interventions and describing comparison conditions, measuring outcomes, and analyzing data. Each QI subsumes a number of specific criteria that a study must meet for it to address the QI. For example, to meet the QI of describing participants, a study must meet these three criteria:

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 characterizing the interventionists or teachers provided? Did it indicate whether they were comparable across conditions? (Gersten et al., p. 152)

Gersten et al. (2005) also proposed eight desirable QIs related to attrition, reliability and data collectors, outcome measures beyond posttest, validity, detailed assessment of implementation fidelity, nature of instruction in comparison condition, audiotape or videotape excerpts regarding the intervention, and presentation of results. In addition to meeting all the essential QIs, high-quality group studies must address at least four of the desirable QIs. Acceptable studies must meet only one of the desirable QIs in addition to addressing all but one of the essential QIs.

The QIs for group studies that Gersten et al. (2005) proposed are somewhat distinct from the criteria for high-quality group research used in other fields. For example, among the study features required for a high-quality group study in special education that the WWC (2008) does not require for a group study that meets evidence standards without reservations in general education are:

- Detailed descriptions of participants, setting, and independent variable, and services provided in the comparison group.
- The use of multiple outcome measures collected at appropriate times.
- Documentation of implementation fidelity
- Appropriate units of analysis (although WWC reviews must note misalignment between units of assignment and units of analysis).

Among the features that the WWC requires for a study that meets evidence standards without reservations but that Gersten et al. does not require for high-quality group studies are overall and differential attrition not severe or accounted for (although Gersten et al. included attrition as a desirable QI), and no intervention contamination. Both sets of criteria for high-quality group studies require researchers to demonstrate the comparability of interventionists across conditions.

Horner et al. (2005) proposed QIs for SSR in special education in seven areas: describing participants and settings, dependent variable, independent variable, baseline, experimental control and internal validity, external validity, and social validity. Horner et al. proposed 21 criteria to assess the presence of these QIs. For example, to meet the dependent variable QI, a study must meet the following criteria:
1. Dependent variables are described with operational precision.

2. Each dependent variable is measured with a procedure that generates a quantifiable index.

3. Measurement of the dependent variable is valid and described with replicable precision.

4. Dependent variables are measured repeatedly over time.

5. Data are collected on the reliability or interobserver agreement (IOA) associated with each dependent variable, and IOA levels must meet minimal standards (e.g., IOA = 80%, Kappa = 60%). (Horner et al., p. 174)

Horner et al. (2005) indicated that reviewers use the QIs, "for determining if a study meets the 'acceptable' methodological rigor needed to be a credible example of SSR" (p. 173). Horner et al. do not explicitly state whether studies must meet all the QIs to be considered of acceptable methodological quality, although we infer that they must. The QIs for high-quality SSR studies in special education overlap somewhat with the criteria for studies that support empirically validated treatments in clinical psychology. Both Chambless et al. (1998) and Horner et al. require that researchers clearly describe participant characteristics and use an appropriate SSR design. Chambless et al. require that researchers compare the intervention with a placebo or another treatment and conduct the intervention by using treatment manuals, whereas Horner et al. do not (although Horner et al. do require that researchers overtly measure fidelity of implementation of the independent variable). Horner et al. require a number of criteria that Chambless et al. do not call for, such as description of physical location, description of the dependent variable with replicable precision, acceptable levels of interobserver agreement regarding the dependent variable, and documentation of the external and social validity of the dependent variable.

**Magnitude of Effect**

For a practice to be considered evidence-based in special education, Gersten et al. (2005) proposed that the weighted effect size of group experimental studies should be significantly greater than zero. We presume that this effect size derives from only those studies found to be acceptable or of high quality vis-à-vis the QIs. For promising practices, Gersten et al. required that a 20% confidence interval for the weighted effect size across studies be greater than zero. In contrast, both clinical psychology (Chambless et al., 1998) and general education (WWC, n.d.b) use statistical significance as the standard to judge whether group studies support well-established treatments and practices with positive effects, respectively. The WWC does consider effect sizes (e.g., $d \geq 0.25$) in the absence of statistically significant findings for determining that a practice has potentially positive effects.

Horner et al. (2005) did not prescribe a particular effect size needed for SSR studies to support a practice. However, for the authors to consider a practice as evidence-based on the basis of SSR in special education, they required a documented causal or functional relationship between use of the practice and change in a socially important dependent variable. Horner et al. suggested that visual analysis "of the level, trend, and variability of performance occurring during baseline and intervention conditions" (p. 171) establishes a functional relationship. Visual inspection of graphic displays of student behavior involves the following:

1. Immediacy of effects following the onset and withdrawal of the practice.
2. Overlap of data points in adjacent phases.
3. Magnitude of change in the dependent variable.
4. Consistency of data patterns across conditions (Horner et al.).

Chambless and Hollon (1998) similarly suggested using visual inspection criteria to determine the effect for SSR studies in clinical psychology. The WWC (2008) is developing guidelines, which are not yet available, for assessing the magnitude of effect in SSR studies.
Gersten et al. (2005) suggested that their proposed criteria and standards for determining EBPs in special education were “merely a first step,” which researchers should refine, “based on field-testing” (p. 163). In response, special education researchers have begun to use the proposed QIs and standards in reviews and analyses of research literature. For example, Browder, Wakeman, Spooner, Ahlgrim-Delzell, and Algozzine (2006) applied the QIs and standards for EBPs proposed by Gersten et al. (2005) and Horner et al. (2005) to 128 intervention studies (88 SSR studies and 40 group quasi-experimental studies) that investigated reading outcomes for individuals with significant cognitive disabilities. Browder et al. condensed the seven QIs and 21 criteria that Horner et al. proposed for SSR into four categories:

- Dependent variable operationally defined and included data on reliability.
- Methods adequately described.
- Data collected on procedural fidelity.
- Baseline and experimental control (with particular focus on, between, and within participant replications).

Two coders independently coded the presence of these four categories for all 88 SSR studies. Interrater agreement was 100% in each category except procedural fidelity, for which interrater agreement was 93%. Fifty-six of the SSR studies met all four of Browder et al.’s (2006) categories of QIs for SSR. From these studies, massed trial as well as systematic prompting met Horner et al.’s (2005) standards for an EBP (i.e., at least five supporting studies involving a minimum of 20 total participants, conducted by at least three different researchers in at least three different locations) for the outcomes of sight-word vocabulary, picture vocabulary, and comprehension. The researchers determined that time delay was also an EBP for sight-word vocabulary and fluency and that pictures were an EBP for comprehension for the target population.

Browder et al. (2006) also clustered the four essential and eight desirable QIs that Gersten et al. (2005) proposed for group research into four categories:

- Outcome measures—operationally defined and evidence of reliability and validity.
- Intervention clearly defined.
- Measure of procedural fidelity.
- Use of comparison group and intervention defined.

Because of the perceived level of judgment required to code these methodological categories, Browder et al. (2006) used a consensus model to establish reliability. In this consensus model, two coders discussed coding decisions until they reached agreement. Therefore, Browder et al. did not report interrater reliability (IRR). Only 2 of the 40 group studies met all four of Browder et al.’s categories for group studies, with no particular practice having sufficient empirical support to be considered evidence-based.

In reviewing the empirical literature on interventions aimed at improving self-advocacy for students with disabilities, Test, Fowler, Brewer, and Wood (2005) assessed the presence of the QIs that Gersten et al. (2005) proposed in 11 group experimental studies and that Horner et al. (2005) proposed in 11 SSR studies. Test et al. found high levels of IRR for coding the QIs in a subset of studies: means of 98.5% agreement for SSR and 98.7% agreement for group experimental research. Test et al. reported that only one of the 11 SSR studies that they reviewed met all the QIs. Although most of the SSR studies met most QIs, only six sufficiently described how participants were selected and only two described and measured procedural fidelity. Test et al. assessed 23 criteria for group studies, examining essential and desirable QIs together and including criteria regarding the conceptualization of a study (that Gersten et al. included in their QIs for research proposals). None of the group studies met all or all but one of Test et al.’s criteria. Among the criteria that few studies met: data collectors unfamiliar with study conditions (n = 4), data collectors unfamiliar with participants (n = 4), documentation of attrition (n = 3), clear descriptions of the difference between intervention and control (n = 3), and measures of procedural fidelity (n = 1). Test et al. did not apply Gersten et al.’s proposed
standards to determine whether any practices evaluated in the reviewed studies were evidence-based.

On a smaller scale, we applied the proposed QIs, as literally as possible, to two group experimental studies (B. G. Cook & Tankersley, 2007) and two SSR studies (Tankersley, Cook, & Cook, 2008). Although the small scope of these pilot projects limits their generalizability, our application of Gersten et al.'s (2005) criteria indicated that the group experimental studies that we reviewed met 40% of the QI components; whereas the SSR studies that we reviewed met 48% of the QI components that Horner et al. (2005) proposed. We reported a moderately low IRR of .69 for SSR QI components (Tankersley et al.; we used a consensus model and did not assess IRR for the group QIs). We found that reliably determining whether studies addressed many of the proposed QIs was difficult because of incomplete and ambiguous reporting in the articles reviewed and because of the lack of specificity and clarity (e.g., operationalized definitions) in the proposed QIs (B. G. Cook & Tankersley; Tankersley et al.).

It is encouraging that both Browder et al. (2006) and Test et al. (2005) applied the proposed QIs and reported high IRR in their coding. However, it is important to note that Browder et al. did not apply all the QIs. Furthermore, Test et al. made no distinction between essential and desirable QIs for group studies and did not apply standards for determining EBPs. Thus, to our knowledge, no published studies have applied the specific QIs and standards for EBPs as Gersten et al. (2005) and Horner et al. (2005) proposed across an entire body of research literature. Clearly, to meaningfully determine the feasibility of applying the proposed QIs and standards and to identify aspects of the QIs and standards that researchers might fruitfully refine, researchers should conduct additional field tests.

**SUMMARY AND ANALYSIS OF FIVE FIELD TESTS OF DETERMINING EBPS IN SPECIAL EDUCATION**

We asked five teams of expert reviewers to faithfully apply the QIs and standards for EBPs, proposed by Gersten et al. (2005) and Horner et al. (2005), to bodies of research literature on interventions relevant to their fields of expertise. Review teams evaluated the intervention literature on five interventions frequently used with students with disabilities: cognitive strategy instruction (Montague & Dietz, 2009); repeated reading (Chard, Ketterlin-Geller, Baker, Doabler, & Apichatabutra, 2009); self-regulated strategy development (Baker, Chard, Ketterlin-Geller, Apichatabutra, & Doabler, 2009); time delay (Browder, Ahlgrim-Delzell, Spooner, Mims, & Baker, 2009); and function-based interventions (Lane, Kalberg, & Shepcaro, 2009). This section summarizes and analyzes the approaches and findings of these five reviews with the goal of making preliminary recommendations for refining the proposed QIs and standards for EBPs, as well as the process for applying them.

**SCOPE OF REVIEW**

Reviewers initially had to determine whether and how to delimit the scope of their review. As Browder et al. (2009) suggests, reviewers might delimit "the specific population of focus, the scope of the dependent variable to be considered . . . , and other aspects of the studies" (p. 360). Four of the five review teams identified a target population more specific than students with disabilities—Baker et al. (2009) and Chard et al. (2009) reviewed studies involving students with and at risk for learning disabilities, Browder et al. focused on students with significant cognitive disabilities, and Lane et al. (2009) targeted students with or at risk for emotional and behavioral disorders. Only Lane et al. included an age parameter, reviewing outcomes for secondary students only. The review teams varied in the degree to which they set parameters for dependent variables. Browder et al. reviewed only studies that specifically assessed picture or word recognition. Montague and Dietz (2009) and Baker et al. stated more general outcome parameters for their reviews—mathematical problem solving and writing performance, respectively. Although Chard et al. and Lane et al. did not specify such outcome variables as inclusion criteria for their reviews, their interventions are associated with particular
outcome areas (i.e., reading for Chard et al. and behavioral outcomes for Lane et al.).

The specific parameters that reviewers apply represent an important concern. Using overly broad parameters (e.g., students with or at risk for disabilities) may not address such critical questions as for whom the practice works with sufficient specificity for research consumers. Conversely, overly narrow parameters may reduce the number of studies available and limit the implications of the review. Although we realize that a variety of sensible rationales exists, for focusing reviews on specific groups, in the absence of a compelling rationale, we recommend that reviews focus on as broad a population as seems reasonable and meaningful and that authors carefully describe participants across studies reviewed to inform consumers about the population for whom the intervention has been shown to be effective.

Determining the Presence of Quality Indicators

A particular element of high-quality research is often neither completely present nor completely absent in a research report but instead is partially present. Recognizing this issue, Baker et al. (2009) and Chard et al. (2009) collaboratively constructed 4-point rubrics for rating the presence of QI components for group experiments and SSR. The other review teams rated each component dichotomously, as met or not met. Because relatively low IRR was associated with using the 4-point rubric, we recommend that future reviews use a dichotomous approach for classifying the presence of QIs, at least until reviewers refine a more detailed rubric that they can use with greater reliability.

Ultimately, the method of choice for identifying the presence of methodological QIs may be a philosophical issue. If the purpose of the reviews is to provide in-depth descriptions of a research base, the use of a rubric—perhaps supplemented with descriptions of the strengths and weaknesses of the literature base for each QI—may be desirable. Alternatively, if the main intent of the reviews is to yield a straightforward decision about whether a practice is evidence-based, the benefit of additional information gained by using a more detailed rating system may not be worth the cost of extra time involved in assessing and reporting the information or the possibility of decreased IRR. Of course, the goals of providing in-depth information on a research base and categorizing practices as evidence-based are not mutually exclusive. Future reviewers in special education may want to provide descriptions, use a multiple-point rating system, and employ a "yes/no checklist" approach, thereby generating reviews to serve different purposes for different audiences.

The Division for Research of the Council for Exceptional Children asked Gersten et al. (2005) and Horner et al. (2005) to identify and briefly describe, not operationally define, sets of QIs (S. L. Odom, personal communication, April 7, 2006) in their development of the QIs. Accordingly, Gersten et al. and Horner et al. stated some of the QIs somewhat subjectively. For example, Horner et al. required that the dependent variable be practical and cost-effective but did not provide concrete guidelines for determining practicality or cost-effectiveness. Accordingly, many of the review teams interpreted, and in some cases modified, the QIs for their reviews. For example, Lane et al. (2009) required that researchers explicitly describe the cost-effectiveness of their intervention. At times, review teams also expanded on the QIs. For instance, Browder et al. (2009) specified that not only must researchers overtly measure implementation fidelity but that they must also document a minimum level of 80%. Lane et al. also required that all components for the internal validity QI for SSR studies be met as a precondition for the external validity QI. In other situations, review teams for this issue reduced the criteria for certain QIs (e.g., Lane et al., 2009, and Montague & Dietz, 2009, set their criteria at 3 data points for baseline, as opposed to the 5 points that Horner et al. suggested). Browder et al. also adapted some of the SSR QIs for the specific outcomes of their review (e.g., they defined the socially important change component as learning at least five new words or pictures).

Browder et al. (2009) suggested that adapting the QIs to optimize their applicability for the intervention being reviewed should be a critical component of each review. Determining whether the QIs can be sufficiently specific and operationalized to yield reliable ratings yet flexible enough to apply meaningfully to a wide variety of
TABLE 1
Summary of Single-Subject Research Quality Indicators Rated as Present

<table>
<thead>
<tr>
<th>Quality Indicator</th>
<th>Participants/setting</th>
<th>Dependent variable</th>
<th>Independent variable</th>
<th>Baseline</th>
<th>Internal validity</th>
<th>External validity</th>
<th>Social validity</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chard, Ketterlin-Geller,</td>
<td>1/6</td>
<td>3/6</td>
<td>2/6</td>
<td>3/6</td>
<td>4/6</td>
<td>0/6</td>
<td>4/6</td>
<td>17/42, 40%</td>
</tr>
<tr>
<td>Baker, Doabler, &amp; Apichatabutra (2009)</td>
<td>1/6</td>
<td>3/6</td>
<td>2/6</td>
<td>3/6</td>
<td>4/6</td>
<td>0/6</td>
<td>4/6</td>
<td>17/42, 40%</td>
</tr>
<tr>
<td>Lane, Kalberg, &amp; Shepcaro (2009)</td>
<td>1/12</td>
<td>5/12</td>
<td>6/12</td>
<td>7/12</td>
<td>2/12</td>
<td>1/12</td>
<td>1/12</td>
<td>23/84, 27%</td>
</tr>
<tr>
<td>Browder, Ahlgrim-Delezell, Spooner, Mims, &amp; Baker (2009)</td>
<td>28/30</td>
<td>30/30</td>
<td>29/30</td>
<td>30/30</td>
<td>29/30</td>
<td>28/30</td>
<td>28/30</td>
<td>200/210, 95%</td>
</tr>
<tr>
<td>Montague &amp; Dietz (2009)</td>
<td>5/5</td>
<td>1/5</td>
<td>5/5</td>
<td>5/5</td>
<td>5/5</td>
<td>5/5</td>
<td>5/5</td>
<td>50/63, 94%</td>
</tr>
<tr>
<td>Total</td>
<td>17/42, 40%</td>
<td>23/84, 27%</td>
<td>200/210, 95%</td>
<td>26/35, 74%</td>
<td>59/63, 94%</td>
<td>325/434, 75%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

studies will be a considerable challenge. Indeed, perhaps some freedom to adapt QIs may be appropriate for certain reviews. We are concerned, however, that giving review teams too much latitude to interpret and adapt QIs may, in some situations, result in reviews that vary considerably in their rigor and findings.

FINDINGS

Quality Indicators. Table 1 summarizes the number of studies meeting Horner et al.'s (2005) QIs for SSR, and Table 2 reports the same information for Gersten et al.'s (2005) QIs for group experimental research (Browder et al., 2009, and Lane et al., 2009, reviewed only SSR studies). The review teams reported widely discrepant findings as to how frequently the studies reviewed met the QIs. The proportion of QIs met in specific reviews ranged from 27% to 95% for SSR studies and from 12.5% to 95% for group experimental studies. It is noteworthy that these considerable disparities were not associated with differences in the rating procedure used. That is, although they all used a dichotomous approach for identifying the presence of SSR QIs, Lane et al. found almost three fourths of QIs absent in the studies that they reviewed, whereas Browder et al. and Montague and Dietz (2009) indicated that almost all QIs were present in the studies that they reviewed. Moreover, both Baker et al. (2009) and Chard et al. (2009) used a 4-point rubric to identify the presence of QIs. However, Chard et al. found that only 25% of the QIs for group experiments were present in the five studies that they reviewed, whereas Baker et al. reported that 95% were present in the five group studies that they reviewed. The disparities in identified QIs may simply reflect significant variation in the methodological quality of the bodies of literature reviewed. Since many of the QIs are not operationally defined, another possibility is that review teams systematically vary in their interpretation of the QIs.

In comparison with the wide discrepancies of QIs met between reviews, the variance in specific QIs present across the studies reviewed was minimal. Of the SSR QIs, the most frequently met was baseline (achieved in 52 of a total of 62 SSR studies reviewed), whereas the least frequently met was independent variable (41 of the 62 SSR studies met this QI). For group experimental studies, the number of total studies that met a QI ranged from 5 (of 12 total studies reviewed) for independent variable/comparison condition to 7 for participants and outcome variable. Across the studies reviewed, the SSR studies met a much higher proportion of QIs than group experiments did. This outcome may have occurred because of differences in the quality of the studies reviewed, differences in the rigor required by the two sets of
QIs, or both. The disproportionately high number of SSR studies reviewed, in comparison with group experiments, may indicate a relative dearth of group experiments in the special education literature (Seethaler & Fuchs, 2005). The small number of group experiments conducted in special education appears to pose a particular concern for those wishing to establish EBPs in the field, given that the results of group experimental research figure prominently in this process.

The identification of components that researchers addressed least often can suggest areas of focus for future researchers to improve the methodological rigor of intervention research in the field of special education. The least frequently addressed component of the dependent variable QI for SSR studies appears to be appropriate documentation of IRR. Issues related to implementation fidelity were clearly the primary reason that studies did not meet the independent variable QI, with each team of reviewers reporting that multiple SSR studies reviewed did not meet this component. For group studies, the least frequently addressed component for the intervention/comparison condition QI was also implementation fidelity. The primary shortcoming of group experiments for the outcome variable QI appears to be not using multiple dependent measures, at least one of which does not tightly align with the independent variable. And the sole reason that group experimental studies reviewed did not meet the data analysis QI was failure to report effect sizes. It is important to note that this special issue reviewed a relatively small number of studies, especially group experimental studies, and that the studies may not represent the larger pool of intervention research in special education, suggesting that these methodological concerns may not be generalizable.

**Interrater Reliability.** Unlike the proportion of QIs met, IRR did appear to vary according to the method used to rate the presence of the QIs. Generally, the three reviews that categorized QIs dichotomously (i.e., present or absent) reported relatively high levels of IRR. For example, Lane et al. (2009) reported 100% IRR for 15 of the 21 SSR components, with only one component falling below 83% (IRR for the component change in dependent variable is socially valid was 75%). Browder et al. (2009) reported a mean IRR across SSR QI components of 97%, with a range from 83% to 100%. And Montague and Dietz (2009) reported a mean IRR of 93% across QIs for SSR studies and 77% for group studies. In contrast, Baker et al. (2009) and Chard et al. (2009)—both of whom used a 4-point rubric to rate the presence of QIs—reported IRR of .36 and 62% for SSR studies and .53 and 77% for group studies. Although IRRs for these two reviews were much higher when allowing for 1-point discrepancies, the reliability for determining the presence of QIs appears to be meaningfully lower when using a 4-point rubric, which is not unexpected, given that Chard et al. reported some difficulties in discriminating between the multiple rating levels. No systematic differences appear to exist for IRR between SSR and group studies. In the three reviews that considered both types of
research, Baker et al. and Chard et al. reported higher IRR for the group studies, whereas Montague and Dietz indicated higher IRR for SSR studies.

**Recommendations for Refining the Process**

On the basis of their experiences applying Gersten et al.'s (2005) and Horner et al.'s (2005) QIs and standards for determining EBPs in special education, the review teams for this issue made a number of recommendations for refining the process. The reviewers suggested adding some new QIs or making some of the existing QIs and their components more rigorous. For example, Chard et al. (2009) proposed requiring researchers to describe the theoretical or conceptual framework for the intervention reviewed (see also Browder et al., 2009). And Montague and Dietz (2009) advocated that researchers specify inclusion and exclusion criteria for selecting participants, assess treatment fidelity with at least two impartial observers with interrater agreement of at least 80%, and report effect sizes for SSR. Moreover, both Chard et al. and Montague and Dietz suggested making some of the desirable QIs for group experiments, such as documenting validity of measurement instruments and minimal attrition, essential QIs.

In contrast to these calls for additional or more rigorous QIs, Lane et al. (2009) suggested that some of the SSR QIs might be overly rigorous. They recommended, for example, that researchers reconsider the requirements for documenting the instruments and process used to determine the disability of participants and describing the cost-effectiveness of the intervention in SSR studies. Lane et al. also advocated that the field consider requiring less than 100% of components for meeting a QI, perhaps using an 80% criterion.

Review teams also noted the need for greater operationalization of the QIs and their components. In particular, Montague and Dietz (2009) called for greater clarity with regard to what constitutes a typical intervention agent in SSR studies. Baker et al. (2009) provided another suggestion for improving the ability of reviewers to determine the presence of QIs in reports of research—furnish opportunities for researchers, perhaps on Web sites linked to the journal, to give additional, detailed information that might otherwise go unreported because of space limitations.

In regard to standards for EBPs, Lane et al. (2009) raised the issue of whether all QIs were equally important, and if not, whether they might be weighted differentially in determining EBPs (see also Montague & Dietz, 2009). Montague and Dietz also suggested that researchers might develop standards for determining when to consider an EBP evidence-based for subpopulations (e.g., how many studies involving students with a particular disability are necessary to demonstrate that the intervention is evidence-based for that population?).

These recommendations all appear to have merit and warrant further consideration while special educators work toward refining the process for determining EBPs in special education. However, we also advise caution in revising the QIs and the process for establishing EBPs too readily or repeatedly. Special educators can and should refine the QIs and standards, perhaps periodically over time, to optimize their efficiency, reliability, and validity. For example, we endorse the idea that the QIs should be further operationalized—a process that the Council for Exceptional Children has undertaken (Bruno, 2007). However, no single set of QIs or standards will meet every purpose; and for the most part, the review teams found the application of the proposed QIs and standards feasible and meaningful. When the QIs and standards have been refined and vetted through what we envision as an iterative but limited sequence of field trials, stability and consistency in the QIs and standards for EBPs in the field will be of significant importance.

**Conclusion**

The authors of the five reviews in this topical issue took on a task that posed multiple challenges. The review teams not only had to systematically review a large number of studies, but they did so by using criteria that often required interpretation while they devised their own processes for field-testing the proposed QIs and standards for EBPs in special education. Not surprisingly,
this process was time-consuming—Browder et al. (2009) estimated that their review team devoted more than 400 hours to their review. The reviewers also no doubt found the review process difficult because we asked them to apply the QIs literally. At times, literally applying the QIs may have seemed to highlight limitations in the research of respected colleagues. It is important to note that authors of previous research wrote the body of extant research without foreknowledge of the future standards of methodological rigor to which it might be held and that they conformed to external requirements of the day (e.g., little emphasis on reporting effect sizes; the perpetual space limitations in journals). Nonetheless, the results of these reviews have provided the first large-scale application of the QIs and the standards for EBPs in special education. We appreciate and applaud the work of the reviewers and the scholars who conducted the original research reviewed, as well as the pioneering work of Cersten et al. (2005) and Horner et al. (2005).

Collectively, the application of QIs to determine high-quality group research and SSR across five bodies of special education intervention research indicate the following:

- Approximately three quarters of the SSR QIs were present across studies reviewed, whereas approximately one half of group experimental QIs were present.
- Considerable variability existed between reviews in the proportion of QIs met. The rating procedure used did not appear to explain this variability.
- The IRR for rating QIs varied markedly between reviews, although reviews using a dichotomous yes/no scheme for identifying QIs tended to yield adequate IRR.

Reviewers also made a number of suggestions for refining the QIs, such as operationalizing them, adding and deleting particular components of some QIs, and weighting the QIs according to their importance. In addition to considering these and other technical matters (e.g., Should reviews be restricted to articles published in peer-reviewed journals?), special education leaders will need to address some foundational issues regarding the need for and merits of determining EBPs in special education so that they can garner the broad support of the special education community for this process.

The philosophical objections to EBPs that we have heard from special educators often parallel criticisms raised regarding the advent of evidence-based medicine. As described by Sackett et al. (1996), “criticism has ranged from evidence based medicine being old hat to it being a dangerous innovation, perpetrated by the arrogant to . . . suppress clinical freedom” (p. 71). Given the documented research-to-practice gap (e.g., B. G. Cook & Schirmer, 2003), the claim that EBPs are old hat seems unwarranted in special education. As for concerns that EBPs in special education will force instruction to conform to an approved menu of interventions, we believe that EBPs will not and should not ever take the place of professional judgment but can be used to inform and enhance the decision making of special education teachers. As Sackett et al. suggested for evidence-based medicine,

Good doctors use both individual clinical expertise and the best available external evidence, and neither alone is enough. Without clinical expertise, practice risks becoming tyrannised by evidence, for even excellent external evidence may be inapplicable to or inappropriate for an individual patient. Without current best evidence, practice risks becoming rapidly out of date, to the detriment of patients. (p. 71)

Likewise, we in no way imagine evidence-based special educators being directed as to when and in what situations they can or cannot use particular teaching practices. Instead, EBPs should interface with the professional wisdom of teachers to maximize the outcomes of students with disabilities (Cook, Tankersley, & Harjusola-Webb, 2008).

EBPs should interface with the professional wisdom of teachers to maximize the outcomes of students with disabilities.

We concur, then, with Sackett et al.’s (1996) declaration that, “clinicians who fear top down cookbooks will find the advocates of evidence
based medicine [or special education] joining them at the barricades" (p. 72). However, although we recognize the dangers of overemphasizing EBPs in a field premised on individualized instruction, we believe that special educators would be remiss if they did not make every effort to prioritize practices shown by our best research to result in meaningful improvements in student outcomes. Identifying practices that are evidence-based for students with disabilities is a necessary but insufficient step in a process that we hope will culminate in the consistent implementation of the most effective practices with fidelity, ultimately resulting in improved outcomes for students with disabilities.

REFERENCES


Exceptional Children


ABOUT THE AUTHORS

BRYAN G. COOK (CEC HI Federation), Professor, Department of Special Education, University of Hawaii, Honolulu. MELODY TANKERSLEY (CEC OH Federation), Professor, Department of Special Education, Kent State University, Kent, Ohio. TIMOTHY J. LANDRUM (CEC VA Federation), Senior Scientist, Department of Curriculum, Instruction, and Special Education, University of Virginia, Charlottesville.

Address correspondence to Bryan G. Cook, University of Hawaii at Manoa, College of Education, Department of Special Education, 1776 University Ave., Wist Hall 117, Honolulu, HI 96822 (e-mail: bgcook@hawaii.edu).

The authors thank the Division for Research of the Council for Exceptional Children for their support of this work and for its leadership in identifying and applying evidence-based practices in special education.

Manuscript received June 2008; accepted September 2008.
Copyright of Exceptional Children is the property of Council for Exceptional Children and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.