CHAPTER 1

Evaluating the Evidence Base for Emotional and Behavioral Disorder Interventions in Schools

Frank M. Gresham and Hill M. Walker

In order to be a sophisticated consumer of approaches to preventing, intervening with, and remediating emotional and behavioral disorders (EBD), professionals need to understand the criteria and standards used to identify and judge approaches that embody acceptable levels of evidence. Among the most popular descriptors used in referring to applied school research over the past several years are "evidence-based treatments" (EBTs) and "evidence-based practices" (EBPs). It is important for educational consumers to understand exactly what the term "evidence-based" means and how it can be used to evaluate any program, assessment procedure, or intervention practice. Professionals often conflate EBTs with EBPs and use the terms interchangeably.

EBTs are particular interventions that have been shown to be efficacious and/or effective through rigorous research methods, most notably the "randomized controlled trial" (RCT). In contrast, EBPs are approaches to intervention rather than specific intervention procedures. In education, EBTs are used to make decisions about individual students (e.g., students may be classified as "responders" or "nonresponders," depending on how they respond to an intervention). EBPs are based on scientific research that supports implementation of certain intervention approaches. A good example of an EBP is the "response-to-intervention" (RTI) paradigm, which is used to change, continue, or terminate an intervention strategy for an individual student through sensitive progress monitoring.

It is important to note that there is not universal agreement about this distinction. For example, in a recent special issue of *Exceptional Children* that builds upon important prior work in school mental health (Burns & Hoagwood, 2002; Hoagwood, Burns, Kiser, Ringeisen, & Schoenwald, 2001; Schoenfeld, 2006) and educational practice (Odom, 2005, 2009), Cook and Odom (2013) define EBPs as programs *and* practices that show meaningful effects on student outcomes achieved through high-quality research from which causality can be inferred. In this paradigm, promising or proven interventions are identified through research that meets rigorous standards and are translated into effective practices through procedures drawn from implementation science. Fixsen, Blasé, Metz, and Van Dyke (2013) provide a formula in which they argue that effective interventions combined with effective implementation equal improved outcomes.

EBTs and EBPs are based on scientific research that supports the use of certain intervention procedures or practices. Evidence for these treatments and practices can be established by using a variety of research strategies. These strategies include carefully summarizing the extant research lit-

erature via meta-analytic methods; conducting experimental and quasi-experimental research studies to support various treatments and practices; analyzing moderators and mediators of various treatments and practices; and conducting tightly controlled single-case experimental design studies. We discuss these strategies further below.

Strength of Evidence and EBTs/EBPs

The research strategies that can be used to marshal evidence (and the strength of the evidence they provide) include, but are not limited to, the following: experimental designs (the strongest evidence), quasi-experimental designs (somewhat weaker evidence), regression discontinuity designs (powerful but seldom used in EBD research), correlation/regression studies (correlational but not causative), single-case experimental designs, quantitative syntheses (meta-analyses), and qualitative syntheses. Numerous syntheses of the evidence literature have attempted to categorize interventions and practices into a false dichotomy of either "evidence-based" or "non-evidence-based." In our view, research evidence does not fall neatly into these two categories, but rather exists on a continuum anchored by evidence-based and non-evidence-based poles. This continuum necessitates thinking in terms of levels or strata of evidence as expressed in categories of stronger or weaker evidence. For example, see Kazdin (2004) for a discussion of the "absolute threshold" versus "hierarchical" approaches to evaluating evidence and judging the strength of applied research. The threshold method is an absolute standard, whereas the hierarchical method is a relative standard that considers a range of evidence generated by differing research methods, in addition to the gold standard of RCTs (e.g., quasi-experimental designs; pre–post outcome studies; correlational studies; descriptive studies using observational methodology; and qualitative, ethnographic and anecdotal evidence). As a rule, we subscribe to the hierarchical approach for establishing evidence as promoted by Kazdin. Ultimately, determining whether a treatment or practice is evidence-based requires evaluating the research methodology used and how well this methodology controls for threats to internal validity, external validity, construct validity, and statistical conclusion validity.

Meta-analyses dating back to the 1970s have shown that a majority of the published intervention procedures for EBD are effective in treating a broad range of externalizing and internalizing behavior problems (Kazdin & Weisz, 2003, 2010). Effect sizes of social-behavioral interventions for children and adolescents often equal or exceed those of widely accepted medical treatments (Ferguson, 2009; McHugh & Barlow, 2012; Rosenthal & Matteo, 2001). However, interventions that have not been subjected to controlled trials are typically considered unproven and/or ineffective. Such interventions cannot be assumed to be either effective or ineffective until they have been rigorously tested and alternative explanations for their achieved effects have been ruled out (see Smolkowski, Strycker, & Seeley, Chapter 31, this volume). Furthermore, Cook and Odom (2013) argue that it is important to distinguish between practices that are not considered evidence-based (1) because they have been shown through a series of high-quality studies to be ineffective, as they do not demonstrate causality; and (2) because an evidence-based review has not been conducted or there is insufficient research evidence to confirm that the practices are effective. There is a consensus among professionals in our field that the most effective interventions, if implemented poorly or incompletely, will not produce acceptable outcomes, and that ineffective interventions, no matter how well implemented, will yield similar results (see Gresham, Chapter 25, this volume).

Types of Research Evidence

The goal of establishing EBPs in our field is to garner the best research evidence related to intervention strategies, types of EBD, and settings in which these interventions are delivered. Multiple types of research evidence can be used to support EBPs; these include (1) efficacy studies, (2) effectiveness studies, (3) cost–benefit/cost-effectiveness investigations, and (4) epidemiological studies. Different types of research designs are better suited to address certain questions than others. These are described below.

- Observation of EBD within target settings, including case studies, can be a valuable source of hypotheses concerning behavioral difficulties of children and youth.
- Qualitative research (see Sabornie & Weiss, Chapter 30, this volume) can be used to describe the subjective or "real-world" experiences of individuals undergoing a particular intervention procedure.
- Single-case experimental designs are useful for drawing causal inferences about the effectiveness of interventions for individuals in a controlled manner (see Smolkowski et al., Chapter 31, this volume).
- Epidemiological research can be used to track the availability, utilization, and acceptance of various intervention procedures.
- Moderator/mediator studies can be used to identify correlates of intervention outcomes and to establish the mechanisms of change in specific intervention procedures.
- RCTs (efficacy studies) provide the strongest type of research evidence and the most protection against various threats to the internal validity of a study (see Smolkowski et al., Chapter 31, this volume).
- A meta-analysis of the research literature provides a quantitative index concerning the effects of multiple studies across various populations, age groups, and settings.

The types of research evidence obtained by using these methodologies can be rank-ordered in terms of their strength based on research design logic. Thus observations can be used to formulate hypotheses, but cannot be used to draw causal inferences about a phenomenon. Single-case experimental designs can be used to draw causal inferences about the effect of an intervention on a given individual, but these effects cannot be generalized to other individuals with somewhat different types of problems. RCTs can be used to draw causal inferences about the efficacy of a given intervention under tightly controlled conditions, but cannot be generalized to other populations, settings, or conditions under less controlled conditions. Quantitative research syntheses (meta-analyses) can provide estimates of the effect sizes of given interventions, but cannot necessarily be used to draw causal inferences about the effects of specific interventions on specific individuals.

Threats to Drawing Valid Inferences

The purpose of research methodology is to design studies uncovering relations among variables that might not be readily apparent from casual observation. Research designs assist in simplifying a complex situation in which many variables are operating concurrently, and in helping researchers to isolate variables of interest. Research designs thus aid researchers in the crucial task of ruling out alternative explanations for the data that are collected in a study. The extent to which any given research design is successful in ruling out plausible rival hypotheses is not absolute, but rather one of degree. In particular, researchers use validity arguments to assist them in ruling out alternative explanations for their data. As noted earlier, four types of validity are typically considered: internal, external, construct, and statistical conclusion (Campbell & Stanley, 1963; Shadish, Cook, & Campbell, 2002). These are described in the following paragraphs.

"Internal validity" refers to the degree to which a researcher can attribute changes in a dependent variable (outcome) to a systematically manipulated independent variable (intervention) while simultaneously ruling out alternative explanations. There are various threats to the internal validity of research studies; these include history, maturation, instrumentation, statistical regression, selection biases, attrition, and interaction of selection biases with other threats to internal validity (see Shadish et al., 2002). The RCT is the gold standard for protecting against virtually all these threats to the internal validity of a research study. Single-case experimental designs also provide protection from many, but not all, of these internal validity threats. Quasi-experimental (nonrandomized studies) designs do not provide this level of protection against internal validity threats.

"External validity" refers to the generalizability of the results of a research study. That is, it asks this key question: To what extent can the results of the study be generalized to

other populations, settings, treatment variables, and measurement variables? The issue of external validity concerns the boundary conditions or limits of research findings. Whereas internal validity is concerned with attributing changes in a dependent variable to an independent variables, external validity is concerned with demonstrating the extent to which the same effect would be obtained with other participants, in other settings, with other treatments, and with different methods of measuring outcomes.

Internal validity is the key concept in "efficacy studies" (investigation of a phenomenon under tightly controlled conditions), whereas external validity is the key feature in "effectiveness studies" (investigation of a phenomenon in "real-world" settings) (Nathan, Stuart, & Dolan, 2000). Several threats to external validity have been identified, and these are classified into four broad categories: sample, stimulus, contextual, and assessment characteristics (Bracht & Glass, 1968).

"Construct validity" refers to the basis for interpreting the causal relation between an independent variable and a dependent variable, whereas internal validity is concerned with whether an independent variable is responsible for change in a dependent variable. Construct validity focuses on the reason for or interpretation of the change in a dependent variable brought about by an independent variable.

The construct validity of a study is based on two questions: What is the intervention? And what explains the causal mechanism for change in the dependent variable? For example, it has been demonstrated that modeling and behavioral rehearsal are two well-established and effective procedures for teaching social skills. The causal mechanism for why these two procedures are effective can be found in research on social learning theory (Bandura, 1977), which has consistently demonstrated that vicarious learning (via modeling) and practice (via behavioral enactment or rehearsal) explain why changes in social skills occur.

"Statistical conclusion validity" refers to threats in drawing valid inferences that result from random error and poor selection of statistical procedures. Statistical conclusion validity deals with those aspects of the statistical evaluation of a study that affect the conclusions drawn from the experimental conditions and their effect on the dependent variable. There are several threats to statistical conclusion validity, including low statistical power (failure to reject a true null hypothesis), unreliability of treatment implementation (poor treatment integrity), unreliability of dependent measures (errors of measurement), random irrelevancies in the experimental setting, and random heterogeneity of respondents.

Levels of Scientific Evidence

Various professional groups have adopted differing but related criteria and nomenclatures for classifying different levels of scientific evidence for interventions. Division 12 (Clinical Psychology), Division 16 (School Psychology), Division 53 (Clinical Child and Adolescent Psychology), and Division 54 (Pediatric Psychology) of the American Psychological Association all have published separate documents specifying criteria for classifying treatments based on the quality of research supporting those treatments. Although there is some variation among these divisions' documents, all have agreed upon what the criteria should be in the classification of scientific evidence. These criteria are described below.

- *Criterion 1: Well-established treatment*. There must be two "good" group design experiments, conducted in at least two independent research settings and by independent research teams, demonstrating efficacy by showing the intervention to be (1) statistically superior to a pill or psychological placebo or to another treatment, *or* (2) equivalent (or not significantly different) to an already established treatment in experiments with sufficient statistical power to detect moderate differences; *and* (3) treatment manuals or their logical equivalent were used for the treatment, conducted with a target population, treated for specific problems, for whom inclusion criteria have been delineated, reliable and valid outcome measures were selected, and appropriate data analyses were used.
- *Criterion 2: Probably efficacious treatment*. There must be at least two good

experiments showing that the treatment is superior (statistically significantly so) to a wait-list control group, *or* one or more good experiments meeting the criteria for well-established treatments with the one exception of having been conducted in two independent research settings and by different investigatory teams.
- *Criterion 3: Possibly efficacious treatment*. There must be at least one "good" study showing the treatment to be efficacious in the absence of conflicting evidence.
- *Criterion 4: Experimental treatment*. The treatment has not yet been tested in trials meeting established criteria for methodology.

Other codifications of standards of evidence, and descriptions of design approaches that produce varying levels of evidence, have been produced by the What Works Clearinghouse of the Institute of Education Sciences and the Society for Prevention Research (see Flay et al., 2005). Glasgow, Vogt, and Boles (1999) have developed a widely cited framework for evaluating the public health impact of health promotion interventions. This framework, called RE-AIM, has five evaluation dimensions:

1. Reach—the proportion of the target population that participated in the intervention.
2. Efficacy—its success rate if implemented according to recommended guidelines and defined as positive outcomes minus negative outcomes.
3. Adoption—the proportion of settings, practices, and plans that will adopt the intervention.
4. Implementation—the extent to which the intervention is implemented as intended in the real world.
5. Maintenance—the extent to which a program is sustained over time.

This RE-AIM framework is directly transferable to the professional subspecialties of school mental health and the field of EBD. Furthermore, it provides a basis for asking searching questions about the nature, efficacy, and effectiveness of approaches commonly used in our field. We encourage professionals to adopt this framework whenever possible in evaluating innovations that are being considered for possible adoption to accomplish prevention and intervention outcomes.

The adoption of interventions and practices for students with EBD in school settings and contexts is increasingly viewed as a consumer protection issue (Detrich, 2008). That is, approaches that are promoted as efficacious or effective need to be accessible and cost-efficient, and must hold the potential to produce acceptable consumer outcomes. "Acceptable" in this instance means that the adopted approach has a reasonable likelihood of solving a problem or remediating a disorder in such a way that (1) the investment of time, effort, and fiscal resources is more than justified by the positive benefits achieved; and (2) participants who are targets and implementers of the approach show high levels of satisfaction based on their exposure to it. We urge professionals to pose two key questions in evaluating the outcomes of an innovation: (1) Is there research evidence that exposure to it moves the participants into or close to the normal range of performance on the outcome measures used? (2) Are outcomes of the intervention and methods used to achieve them acceptable to target consumers (parents, students, teachers)?

Numerous lists and inventories of recommended interventions and approaches are now broadly available, but many have not been thoroughly vetted against the four criteria described above, codified evidence standards (Cook & Odom, 2013), the RE-AIM framework, or the approaches and evaluative guidelines described by Smolkowski and colleagues in Chapter 31 of this volume. We believe that the measures, interventions, and practices reviewed and recommended by contributors to this handbook provide a basis for judging whether they can be considered promising and/or proven and meet the standards of acceptable evidence. Practicing professionals who adopt and implement them can have reasonable confidence that they will work as described, provided that they are implemented with high levels of treatment integrity and that obstacles to such implementation are systematically addressed.

We are hopeful that our field will adopt a science of educational and school-related

EBD research within the next decade along the lines so well described by Kauffman herein. If this occurs, we believe that many of the current school-based barriers to the adoption of effective practices for the student population with EBD are likely to be reduced and attenuated.

References

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bracht, G., & Glass, G. (1968). The external validity of experiments. *American Educational Research Journal, 5*, 437–474.

Burns, B., & Hoagwood, K. (2002). *Community treatment for youth: Evidence-based interventions for severe emotional and behavioral disorders*. New York: Oxford University Press.

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

Cook, B., & Odom, S. (2013). Evidence-based practices and implementation science in special education. *Exceptional Children, 79*(2), 135–144.

Detrich, R. (2008). Evidence-based, empirically supported, or best practice?: A guide for the scientist-practitioner. In J. K. Luiselli, D. C. Russo, W. P. Christian, & S. M. Wilczynski (Eds.), *Effective practices for children with autism* (pp. 3–25). Oxford, UK: Oxford University Press.

Ferguson, C. (2009). Is psychological research really as good as medical research?: Effect size comparisons between psychology and medicine. *Review of General Psychology, 33*, 130–136.

Fixsen, D., Blasé, K., Metz, A., & Van Dyke, M. (2013). Statewide implementation of evidence-based programs. *Exceptional Children, 79*(2), 213–233.

Flay, B., Biglan, A., Boruch, R., Castro, F., Gottfredson, D., Kellam, S., . . . Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science, 6*(3), 151–175.

Glasgow, R., Vogt, T., & Boles, S. (1999). Evaluating the public health impact of health promotion interventions: The RE-AIM framework. *American Journal of Public Health, 89*(9), 1322–1327.

Hoagwood, K., Burns, B., Kiser, L., Ringeisen, H., & Schoenwald, S. (2001). Evidence-based practice in child and adolescent mental health services. *Psychiatric Services, 52*, 1179–1189.

Kazdin, A. E. (2004). Evidence-based treatments: Challenges and priorities for practice and research. *Child and Adolescent Psychiatric Clinics of North America, 13*(4), 923–940.

Kazdin, A. E., & Weisz, J. R. (Eds.). (2003). *Evidence-based psychotherapies for children and adolescents*. New York: Guilford Press.

Kazdin, A. E., & Weisz, J. R. (Eds.). (2010). *Evidence-based psychotherapies for children and adolescents* (2nd ed.). New York: Guilford Press.

McHugh, R., & Barlow, D. (Eds.). (2012). *Design and implementation of evidence-based interventions*. New York: Oxford University Press.

Nathan, P., Stuart, S., & Dolan, S. (2000). Research on psychotherapy efficacy and effectiveness: Between Scylla and Charybdis? In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (3rd ed., pp. 505–546). Washington, DC: American Psychological Association.

Odom, S. L. (Ed.). (2005). Criteria for evidence-based practice in special education [Special issue]. *Exceptional Children, 71*(2).

Odom, S. L. (2009). The ties that bind: Evidence-based practices, implementation science, and outcomes for children. *Topics in Early Childhood Special Education, 29*, 53–61.

Rosenthal, R., & Matteo, R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59–82.

Schoenfeld, A. H. (2006). Design experiments. In J. L. Green, G. Camilli, P. B. Ellmore, & A. Skukauskaite (Eds.), *Handbook of complementary methods in education research* (pp. 193–206). Washington, DC: American Educational Research Association.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.