TASK FORCE ON EVIDENCE-BASED INTERVENTIONS IN SCHOOL PSYCHOLOGY

Sponsored by: Division 16 of the American Psychological Association

and

The Society for the Study of School Psychology

Endorsed by: The National Association of School Psychologists

For further information contact:

Thomas R. Kratochwill, PhD
Chair, Task Force on Evidence-Based Interventions in School Psychology
School Psychology Program
1025 W. Johnson Street
University of Wisconsin-Madison
Madison, WI 53706
.tomkat@education.wisc.edu
(608) 262-5912
Table of Contents

Preface 6
Task Force on Evidence-Based Interventions in School Psychology 6
Future Considerations 7

Procedural and Coding Manual for Review of Evidence-Based Interventions 9

Section 1: General Considerations 9

Procedures for Identifying Relevant Intervention Outcome Studies 9
Literature Reviewers 9
Task Force Domains 10
School- and community-based intervention programs for social and behavioral problems 10
Academic intervention programs 10
Family and parent intervention programs 10
School-wide and classroom-based programs 10
Comprehensive and coordinated school health care services 10

Literature Search Process 11
Acceptable Sources in the Literature: Peer Review Required 11
Coding Qualitative Research Methods and Mixed Qualitative-Quantitative Designs 10
Coding Mixed Group and Participant Designs 10
Include Refuting Evidence and Null Findings 11
Obtaining Missing Information 12
Limits of Historical Search 12
The Literature Review Process 12
Figure 1. Framework for the Literature Review Process 13

Section 2: Criteria for Coding Studies 15

Section 3: Coding Manuals and Technical Coding Protocols 18

Coding Manual for Group-Based Design Intervention Research 22

Group-Based Designs 22
General Characteristics 22
Key Features for Coding Studies and Rating Level of Evidence/Support 30

Coding Protocol: Group-Based Designs 52

Coding Manual for Single-Participant Intervention Research 63

Single Participant Designs 66
General Characteristics 66
Key Features for Coding Studies and Rating Evidence 70
Other Descriptive/Supplemental Criteria 79
Coding Protocol: Single-Participant Design 84

References 114
## Appendixes

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appendix A</td>
<td>Former Members and Domains of the Task Force on Evidence-Based Interventions in School Psychology</td>
<td>118</td>
</tr>
<tr>
<td>Appendix B</td>
<td>List of Task Force Members and Content Information</td>
<td>119</td>
</tr>
<tr>
<td>Appendix C</td>
<td>General Guidelines for Conducting a Literature Review</td>
<td>125</td>
</tr>
<tr>
<td>Appendix D</td>
<td>General Resources</td>
<td>131</td>
</tr>
<tr>
<td>Appendix E</td>
<td>Supplemental Coding Options for Quasi-Experimental Group-Based Designs</td>
<td>134</td>
</tr>
</tbody>
</table>
Preface

Task Force on Evidence-Based Interventions in School Psychology

The Task Force on Evidence-Based Interventions in School Psychology (hereafter called the Task Force) was formed in 1998 and operated through 2007. It was supported by the American Psychological Association (Division 16 – School Psychology) and the Society for the Study of School Psychology (SSSP) and endorsed by the National Association of School Psychologists (NASP). A detailed history of the Task Force and its focus are presented in several publications [see Kratochwill & Stoiber, 2000, 2002; Stoiber & Kratochwill, 2000; and the Mini-Series in School Psychology Quarterly, 17(4)]. Past membership and contact information for the Task Force is included in Appendix A.

The Procedural and Coding Manual for Review of Evidence-Based Interventions (2nd ed.) (hereafter called the Manual) was formed to identify, review, and code studies of psychological and educational interventions for behavioral, emotional, and academic problems and disorders for school-aged children and their families. The Manual also focuses on interventions that promote health and wellness as well as prevention programs for a variety of areas. Specifically, the Manual is intended to assist in reviewing outcome research for the following purposes:

a) to identify prevention and intervention outcome studies bearing on the effects of educational and psychological prevention and intervention programs,
b) to code those studies according to Task Force criteria, while providing information on the characteristics of the interventions and the studies testing their effects,
c) to determine to what degree the interventions are evidence-based on a variety of criteria,
d) to offer the field of school psychology and related fields some guidelines for using and adopting effective programs, and
e) to provide a template for improving research in the field of psychology and education.

The primary purpose of the document is to help professionals identify, review, and code interventions that have been subjected to empirical research and evaluation. In development of the Manual we have been fortunate to take into account the concepts advanced through various drafts of the Procedural and Coding Manual for Identification of Beneficial Interventions produced by the Committee on Science and Practice by the Society for Clinical Psychology Division 12 of the American Psychological Association (Weisz & Hawley, 2002). In some sections of the Manual we have used material with permission of the Division 12 Committee. The Manual also takes into account the Criteria for Evaluating Treatment Guidelines (American Psychological Association, 2000) produced by the Template Implementation Work Group of the Board of Professional Affairs, Board of Scientific Affairs, and the Committee for the Advancement of Professional Psychology.

The Manual reflects some specific considerations about applying evidence-based interventions (EBIs) in school contexts as articulated elsewhere (e.g., Kratochwill & Stoiber, 2000, 2002; Stoiber & Kratochwill, 2000) and by Task Force committee members. We envision the identification of EBIs to be a long-term, ongoing process. As technologies develop and empirical evidence about an intervention accumulates, the status of the interventions in the EBI database will also evolve. The intent of the Manual is to assist the profession in developing a meaningful database on EBIs.

A major focus of the Task Force has been the development of the Manual, described in more detail below. The Manual represents an effort on the part of many individuals. We express sincere appreciation to Karen O’Connell, Task Force support person at the University of Wisconsin-Madison, for her work on numerous drafts of this document.

Similar to our Division 12 predecessors who have undertaken the task of developing a coding manual, we view this Manual as a document in progress. Knowledge and issues bearing on identification and review of EBIs are
constantly changing and the Task Force has updated the Manual presented here. The Manual has been organized into three sections. The first section describes procedures for identifying intervention studies that are appropriate for review in the domains established by the Task Force. The second section provides specific details regarding Task Force criteria for coding intervention outcome studies and rating studies on a number of predetermined criteria. The final section of the Manual includes technical codes, and is intended to provide users with a detailed understanding of how to code the evidence bearing on various interventions. This section of the Manual is the most detailed and is organized into two subcomponents (called Submanuals) that feature different methodological approaches to reviewing evidence from intervention studies. Each Submanual in this section provides an overview of the coding criteria and is accompanied by a technical coding protocol. Specifically, the following Submanuals are featured:

**Group-Based Designs.** These studies involve traditional multiple-participant investigations that classify research designs based on definitions presented by Kirk (1995). These designs include completely randomized designs and randomized block designs (both between- and within-subjects variations). The design types are applied to studies where random assignment of experimental units to intervention either is or is not included.

**Single-Participant Designs.** These studies involve traditional single-participant investigations that classify research designs based on definitions presented by Hayes, Barlow, and Nelson-Gray (1999). These designs include within-series, between-series, and combined-series and other design variations. Designs can also be coded whether randomization is used.

**Future Considerations**

The criteria presented in the Submanuals are designed to help individuals make decisions about the quality of research evidence and empirical support of various prevention and intervention programs implemented in school and applied settings. Specifically, the Task Force was interested in making state-of-the-art conclusions pertaining to research support for various intervention programs across five domains identified later in this document. However, the Task Force also emphasizes that one of the primary purposes for identifying interventions with empirical support is so that practitioners in schools and other applied settings use these interventions. Yet, generalization from research to practice settings is not a straightforward process and often a variety of barriers exist in this process (Kazdin, 2008; Kazdin, Kratochwill, & VandenBos, 1986; Kratochwill & Stoiber, 2000, 2002; Stoiber & Kratochwill, 2000).

Among the variety of tactics that might be invoked to deal with the hiatus that exists between research and practice, several have been considered by the Task Force to facilitate improved generalization (Hayes et al., 1999; Kazdin et al., 1986). First, standardization of the assessment and intervention process is often associated with work on evidence-based interventions and the Task Force recommends that individuals use manuals or other procedural guidelines to facilitate the process of implementing evidence-based interventions in applied settings. Second, the Task Force recommends mechanisms for training to reduce the hiatus between research and school applied activities. One of the major goals of the Task Force was to create a database that will be useful to training programs and professional organizations in educating practitioners, graduate students, university trainers, etc., in tactics to facilitate the dissemination of research findings to practice settings for various interventions that appear to have strong empirical support. Third, the Task Force recommends that practitioners embrace a scientist-practitioner model in their professional work as a tactic to reduce the hiatus between research and practice. Despite the best efforts of the Task Force to address various issues surrounding intervention implementation in research, it is quite apparent that no series of research investigations can take into account all contextual and ecological variables that have an influence on whether an intervention is likely to be beneficial in a particular setting such as a school, classroom, etc. Therefore, it is essential that practitioners engage in self-evaluation and apply methods of empirical clinical practice to help the field understand how beneficial an EBI will be when applied in a new school context or different situation (Kratochwill & Shernoff, 2003).
The Task Force also hopes that the process of clinical replication will be invoked where practitioners can evaluate various interventions under conditions in which they are likely to be practiced (e.g., effectiveness and transportability studies conducted under real world conditions). Single-participant case study investigations, and related methods can be used to facilitate not only an evaluation of an intervention within a particular context, but also to produce information related to communication among researchers about how to revise or redesign intervention programs to make them more effective in practice settings (Hayes et al., 1999). Although a lofty goal, the Task Force hoped to promote the use of research findings in practice and during the latter stages of our activities, we hope to embrace this important training agenda (see Kratochwill & Stoiber, 2002). The Task Force formed a Research-to-Practice subcommittee in 2003 that will address this important agenda. They are in the process of reporting out their results.

Finally, the Task Force wishes to emphasize the importance of taking into account multicultural and diversity issues in development of the Manual. Intervention studies with high levels of external validity and cultural validity offer information critical for determining the potential for transferability and use with differing populations (Ingraham & Oka, 2006). Sue (1999) called for greater attention to external validity and for researchers to go beyond the use of demographic or proxy variables (such as race, age, gender, ethnicity), to identify various cultural influences and to identify the deeper psychological experiences and interpretations of cultural membership (e.g., acculturation, perceptions of power, oppression) that characterize particular individuals and groups. Culture is defined as the shared norms, values, beliefs, practices, and behaviors relevant to the target population; and reflect cultural experiences that extend beyond the traditional categories of race, ethnicity, and language to encompass the shared experiences of any group (e.g., culture of classroom, school, neighborhood, family, peer group). Wampold (2002) argued that therapists’ effects accounted for nine times more variance than treatment effects. Building on Quintana’s ideas, Ingraham and Oka (2006) provide a a framework for assessing the cultural validity of research. Thus, research on interventions needs to carefully describe the participants, the interventionists, and the methods used to assure that both intervention and methods for documenting evidence and evaluating intervention effectiveness (e.g., research questions, methods, and interpretations) are consistent with the culture(s) of the participants.

Quintana, Troyano, and Taylor (2001) proposed the addition of cultural validity to existing types of research validity: Internal, External, Construct, Hypotheses and Statistical Conclusions. Quintana et al., (2001) advocate increasing the validity of research by critically evaluating the cultural validity of the study’s design, procedures, interpretation, and discussion of results. Thus, cultural validity features are included within this revision of the Manual. Intervention studies will be evaluated to the extent that they provide deeper descriptions of the cultures and values of those involved with the research, including the participants, collectors of data, and researchers. In some sections, the coding criteria reflect a level of attention to cultural and contextual factors that is aspirational, but we believe these are critical in understanding the extent to which interventions have specific or more universal effectiveness. We have attempted to be inclusive in these considerations not solely with regard to the nature of the target clientele, the interveners, and the researchers, but also with regard to research methods that include qualitative, quantitative, and mixed methods. We are requesting reviewers of the empirical literature to be especially sensitive to cultural diversity and issues surrounding the context of intervention implementation in schools and community settings.

With the help of our Task Force cultural diversity domain co-chairs Colette Ingraham and Evelyn Oka, Bonnie Nastasi, and the advice of Division 17 Task Force member Steve Quintana, we developed criteria that consider cultural and ethical issues in the evaluation of intervention research. While many of these criteria are aspirational, the Procedural and Coding Manual also provides guidelines for the design of future research. In the present version of the Manual, the criteria that consider cultural and ethical issues in the evaluation of research have been embedded within the Coding Manual for Group-Based Design Intervention Research and within the Coding Manual for Single-Participant Intervention Research.
PROCEDURAL AND CODING MANUAL FOR REVIEW OF EVIDENCE-BASED INTERVENTIONS

SECTION 1: General Considerations

Procedures for Identifying Relevant Intervention Outcome Studies

The general goal of the Task Force is to promote reviews of the research literature on a particular intervention. Several procedures are required to identify evidence-based interventions. First, a thorough search for relevant evidence in the research literature is required. Such literature might range from a single published article to a collection of studies bearing on a particular intervention. Intervention programs and techniques can be implemented in schools and/or community settings. The program or technique should have relevance to school psychology and may or may not involve clinical populations. Appendix C provides some general guidelines for conducting a literature review as well as an example literature search. Further information is available in several sources (e.g., Cooper, 1998; Fink, 1998; Hart, 2001; Lipsey & Wilson, 2001; Pan, 2008). Appendix D also provides a listing of additional resources.

Literature Reviewers

The reviewers who carry out the literature search have been members of the Task Force or others working independently. The Task Force focused primarily on literature review content that was organized into five domains. Initially, the Task Force developed a separate prevention domain. However, after some discussion a decision was made to identify evidence-based prevention programs within each of our five content domains.

Prevention programs focus primarily on the promotion of social, emotional, and behavioral competencies. Prevention programs reduce the incidence of mental health problems and are implemented before a problem becomes salient or is diagnosed as a disorder. Thus, all children, as well as children who are at-risk for social, emotional, and behavior dysfunction, are targeted. Although effective prevention programs have existed for some time, school-based prevention programs for issues such as substance abuse, violence, or other risk behaviors have become more prevalent in recent years. Programs in this domain generally include a prevention-focused curriculum that is implemented alone or as part of a more comprehensive school-wide or classroom-based program.

The decision to infuse prevention programs into each of the domains was based on the following considerations. First, the Task Force wanted an organizational framework that would allow identification of EBIs by content domain. Second, the Task Force found that prevention programs often address a continuum of intervention impact points ranging from at risk and protective factors to problems experienced by children (e.g., Minke & Bear, 2000). Third, prevention programs for one target area may be intervention programs in another domain. Thus, our decision was to include a wide range of prevention/intervention programs within our domains, and to classify them as universal, selective, or indicated programs.

Task Force Domains

The following five domains were established by the Task Force:

School- and community-based intervention programs for social and behavioral problems. A wide variety of intervention programs for social/behavioral problems and disorders have been implemented in school-based settings (see Kazdin & Weisz, 2003; Morris & Kratochwill, 1998; Stoiber & Kratochwill, 1998). These intervention programs are implemented once a problem or disorder has been detected. The children that participate in these interventions usually share an essential mental health issue, circumstance, or psychological problem, such as being an abuse victim, suffering from depression, lacking anger-control, or being a child in a divorced family. In
their attempt at responding to a wide range of student needs, school psychologists may implement prevention and/or intervention at the individual, group, classroom, or school-wide level.

**Academic intervention programs.** School and educational psychologists have been involved for many years in implementing intervention programs for a wide variety of academic problems (e.g., Rathvon, 1999; Shapiro, 1996). The domain will feature reviews of academic programs for the major academic areas of the curriculum (e.g., reading/literacy, mathematics, written language). Academic programs for children identified with exceptional education needs and those considered at risk within the general education curriculum will be a primary focus within this domain. Some examples of academic intervention programs include those in the area of reading (word recognition and related processes, fluency, comprehension), writing (handwriting, spelling, composition), and mathematics (computation and related processes, concepts and problem solving). Similar to the intervention programs for social and behavioral problems, consultative technologies would be appropriate for this domain.

**Family and parent intervention programs.** Family intervention programs are becoming increasingly important for school psychologists and related professionals working with children and adolescents (e.g., Christenson & Conoley, 1992; Fine & Carlson, 1992). The purposes of reviews in this area will be both family-based intervention and home-school collaboration projects that are implemented in schools or coordinated with school settings, and demonstrate a change in the child’s behavior/performance in school. Because family involvement is a significant, positive correlate of students’ academic achievement and social adjustment, schools have developed specific efforts and practices for intervening with families and for promoting home-school collaboration to enhance student success. Intervention subdomains include parent education, preschool interventions (early childhood/ family education), parent involvement, family/parent consultation, family-school partnership, and family systems.

**School-wide and classroom-based programs.** Fueled by the demands to address increasing social and behavioral problems among youth, school-wide or whole classroom programs are being advocated to provide more integrative solutions for children (Brophy, 2000). Recent literature emphasizes the critical combination of modifying teacher behavior, classroom environment, and school climate as well as actively teaching prosocial behaviors across the curriculum (e.g., Perry & Weinstein, 1998; Roese, Eccles, & Strobel, 1998). Some examples of programs included in this domain are those addressing classroom management, system reform, inclusion, and school collaboration. Interventions such as these, which can be either preventive or remedial, typify organizational programs that would be appropriate for this domain.

**Comprehensive and coordinated school health services.** In recognition of the interrelationships among physical, psychological, social and educational functioning, school psychologists are being called upon to address the physical and mental health needs of students through participation in comprehensive and coordinated school health services (Adelman & Taylor, 1998; Power, Healthfield, McGoey, & Blum, 1999; Short & Talley, 1997). For the purposes of this review, comprehensive and coordinated school health services includes school-based or school-linked services “related to a broad spectrum of health-related problems, including chronic medical or health-related conditions (e.g., endocrine and seizure disorders, childhood cancer, asthma and allergies, diabetes, HIV/AIDS, fetal alcohol syndrome; School Psychology Review, Volumes 24 [2] & 28 [2]), psychiatric disorders (e.g., mood and behavioral disorders; Doll, 1996), and social morbidities (e.g., drug abuse, sexually-transmitted diseases, suicide, violence; DiClemente, Hansen, & Ponton, 1996).” Authors will review research related to comprehensive health care interventions in the school health areas adopted by the Centers for Disease Control and Prevention: Health education; physical education; health services; nutrition services; healthy school environments; health promotion for staff; and parent/community involvement. To the extent these areas overlap with previously defined sections (e.g., programs for social-behavioral problems, school-wide and classroom-based), they will be treated in their respective sections. The emphasis in this domain is on primarily health-related concerns. Current models of comprehensive health care suggest the need for attention to several key factors: (a) integration of educational, health and social services; (b) interdisciplinary and inter-agency collaboration; (c) attention to ecological, contextual, and individual factors; (d) developmentally and culturally appropriate services; (e) a continuum of care ranging from prevention to treatment; (f) provision of empirically supported care; and (g) routine and systematic program evaluation (e.g.,
Adelman & Taylor, 1998; Dryfoos, 1995; Kolbe, Collins, & Cortese, 1997; Roberts & Hinton-Nelson, 1996; Nastasi, Moore, & Varjas, 2004). Although evaluation research is generally supportive of comprehensive health services for children and adolescents (e.g., Dryfoos, 1995; Jordan, 1996; Saxe, Cross, Lovaas, & Gardner, 1995), findings are not unambiguous. Furthermore, given the community-based nature of much of the current work, attention specifically to school-based and school-linked service provision is warranted.

**Literature Search Process**

Reviewers can search for relevant studies using, at a minimum, the following procedures: (a) computerized data bases (e.g., PsycInfo); (b) published reviews and meta-analyses relevant to the target domain; (c) hand searches of the most relevant journals (as determined by the reviewers) to identify appropriate studies not detected through computer search procedures; (d) studies suggested by authors of the interventions being considered; (e) studies suggested by other investigators or practitioners in the field, and (f) reference trails generated by studies thus identified. Review articles and meta-analyses serve the purpose of identifying studies to be reviewed as a starting point; the original articles on which the reviews are based must be reviewed to ensure that the results have been correctly described and appropriately interpreted with criteria invoked by the Task Force. Readers should consult Appendix C for further information on conducting literature reviews.

**Acceptable Sources in the Literature: Peer Review Required**

To insure some level of quality control, a consistent requirement for acceptability of the studies identified is that each must have been subjected to peer review. In most cases, this criterion will mean that the studies have been published in peer-reviewed journals; in some cases, intervention outcome studies published in books will be acceptable if there is evidence that critical peer review occurred. The Task Force does not recommend review of dissertations in their unpublished form. The peer review requirement will mean that the following will NOT be considered acceptable evidence by the Task Force: unpublished manuscripts, book chapters, and information on the Internet that has not been subjected to the peer review (except in the rare cases where critical peer review is evident).

**Coding Qualitative Research Methods and Mixed Qualitative-Quantitative Designs**

Specific criteria were developed by the Qualitative Research Committee for evaluating the use of qualitative research methods in intervention studies. A full description of the criteria specific to qualitative research can be found in Nastasi and Schensul (2005), introduction to the special issue of *Journal of School Psychology* (Volume 43, No. 3) devoted to qualitative research methods in intervention studies.

Recognizing the applicability of the criteria for all intervention studies and for assessing cultural validity, the Task Force chose to integrate criteria related to qualitative and quantitative research into one set of criteria for respective group and single participant designs. This integration reflects the perspective that all intervention studies should be evaluated with the full set of criteria included herein. In addition, studies that incorporate both qualitative and quantitative methods can be adequately coded with this version of the coding manuals.

**Coding Mixed Group and Participant Designs**

In some cases reviewers may be faced with the task of coding intervention research that is based on more than one methodology featured in this manual (both group and single-participant designs). In such cases, the reviewers have the option that (a) either one primary methodology is coded, with the secondary methodology coded on relevant criteria, or (b) that both methodologies be fully coded.
Include Refuting Evidence and Null Findings

Relevant studies include not only those showing beneficial effects of the tested intervention, but also any study showing that the intervention has harmful or null effects (Kratochwill, Stoiber, & Gutkin, 2001). These characteristics will be noted when reviewers evaluate the individual studies.

Obtaining Missing Information

When data necessary for classifying interventions are not provided in the published studies (e.g., means and standard deviations), and there are no other data from which to calculate effect sizes, reviewers are provided with guidance in the Manual regarding how to calculate effect sizes. Reviewers will contact study author(s) and request the missing data when additional information is needed to code the study. Until and unless author(s) provide the necessary data, those studies will not be considered as supporting evidence for the interventions described therein.

Limits of Historical Search

Typically, all research studies in existence bearing on a prevention or an intervention domain will be included in the review. Studies dating back 25 years or more will typically not be included unless the study is of special significance.

The Literature Review Process

Generally, reviewers will follow a structured process for conducting the literature review. Figure 1 presents the general framework for this process.
Figure 1. Framework for the Literature Review Process.
SECTION 2: Criteria for Coding Studies

Once outcome studies relevant to a particular intervention have been identified, the studies will undergo review and a rating will be made about the type of evidence that is available for each prevention/intervention program. In this first section of the Manual the Task Force lists criteria for reviewing the evidence in support of intended intervention effects and provides some operational definitions of various terms. More specifically, the Task Force expands the Division 12 criteria based on additional criteria presented by Crane (1998) and Troia (1999) in their literature reviews and Lipsey and Wilson (2001) in their discussion of meta-analysis.

Because of potential problems in relying on experimental design features to determine whether an intervention produced the intended effects on the participants who are presumed to benefit (i.e., children, families, teachers, etc.), the method the Task Force has chosen for evaluation is delineated into three types: (a) general characteristics, (b) key features, and (c) supplemental descriptive information. The first type of coding consideration in the Manual addresses the general methodological qualities and statistical procedures used in determining the effects of an intervention. These criteria refer to the credibility of the intervention approach and the quality of the outcome evaluation. Reviewers will take into account the overall methodological quality of the outcome evaluation, and in particular, the use of appropriate evaluation methods that correspond to features of the school environment. This first set of information should be helpful in providing a context for understanding the research conducted to demonstrate intervention effects (for example, whether the study incorporated a group or single-participant design is noted).

Twelve key features of an intervention study will be rated on a 4-point scale to specify level of evidence (i.e., 3 = strong evidence/support, 2 = promising evidence/support, 1 = marginal or weak evidence/support, 0 = no evidence/support). These key features focus both on internal and external validity criteria, as well as on features considered important for school- or field-based implementation. The 12 key features include: (a) Research Methodology, (b) Measurement, (c) Comparison Group, (d) Primary/Secondary Outcomes Significant, (e) Cultural Validity or Significance, (f) Educational/Clinical Significance, (g) External validity, (h) Durability of Effects, (i) Identifiable Intervention Components, (j) Implementation Fidelity, (k) Replication, and (l) School- or Field-based Site (see American Institutes for Research & Herman, 1999, An Educators’ Guide to Schoolwide Reform for an example of this type of rating).

Generally, a premium is placed on studies incorporating a comparison group that permits judgment of the merits of an intervention. Because experience with school-based or field-based investigations has often revealed practical problems in implementing random assignment (Lipsey & Cordray, 2000), consideration will be given to alternative methodological and statistical strategies incorporated to contend with internal validity threats. In addition, reviewers will take into account various internal validity issues that might influence interpretation of outcomes even when random assignment has been used (e.g., resentful demoralization, differential attrition). See Shadish, Cook, and Campbell (2002) and Levin and O’Donnell (1999) for further information on this issue. Another type of information that will be made available to readers describes important qualities of the study related to external validity. This type of information is particularly useful for individuals who are interested in evaluating whether the intervention is matched well to specific program needs in applied settings. These features also are reflected in the criteria as Identifiable Intervention Components.
SECTION 3: Coding Manuals and Technical Coding Protocols
CODING MANUAL FOR

GROUP-BASED INTERVENTION RESEARCH DESIGN

Task Force on Evidence-Based Interventions in

School Psychology
GROUP- BASED DESIGN

I. General Characteristics

This first area considers the general characteristics of the study itself, including the conceptual foundations, design, treatment of data, and type of intervention involved. These considerations help readers to understand the paradigm and the approach being used to conduct the intervention research in a particular investigation. Together, the items in this first section create the foundation from which the specific key features of the study can be examined and evaluated. As such, any ratings of these characteristics are intended to be descriptive in nature and may or may not be linked to the study’s overall level of evidence regarding efficacy or effectiveness of the intervention.

A. General study characteristics addresses the foundations for the study and the logical connections among the theoretical-empirical basis, research questions, and methodology. With regard to cultural diversity considerations, this section addresses the rationale and appropriateness of the study’s conceptual basis, questions, and methods in terms of the target population. Issues to consider include:

- The extent to which the theoretical basis and key constructs for the study were grounded in etic (researcher) and emic (participant) perspectives; for example, is the study based on an integration of both existing theory and formative research to test the relevance of existing theory to the cultural ecology and experiences of the target populations.
- The rationale for the theoretical base and key constructs was explicit; that is, the link of researcher’s and participants’/stakeholders’ perspectives to the conceptual basis of the study were clearly stated.
- The research questions reflect the views of the researcher and participants (etic-emic views) and guide the selection of methodology and sample.

A1. Theoretical basis. Intervention design is based upon a conceptual model that is grounded in theory and applied to the empirical study of the target phenomenon and group. The rating on this item reflects the extent to which the theoretical basis is described sufficiently and guides (is linked to) the intervention.

☐ = Described theoretical basis and clearly linked to the intervention.
☐ = Described theoretical basis but no connection to the intervention.
☐ = Provided some reference to theory but lacks clear conceptual model and its relation to the intervention.
☐ = Theoretical basis absent.

A2. Ecological validity. This characteristic refers to extent to which the relevance or appropriateness of the study’s empirical-theoretical (conceptual) basis has been established for the target population. To receive the maximum rating, ecological validity must be established through formative research that is conducted with the target population and used to inform the design of the intervention with the target population to (a) examine the relevance of the conceptual model to the specific population and determine the need for adaptation, or (b) develop a culture/population-specific conceptual model. In the absence of formative research, existing theory could be adapted based on existing research with the target population or similar/related populations; these alternatives are reflected in lower ratings.

☐ = Ecological validity of conceptual base established through formative research with target population OR conceptual base is culturally derived.
☐ = Existing theory adapted based on research with target population.
☐ = Existing theory adapted based on empirical evidence with similar or related populations.
☐ = Existing theory used without attention to ecological validity.

A3. Researcher perspective. This criterion refers to the extent to which the researcher has articulated personal beliefs related to the conceptualization or empirical-theoretical basis of the study.

☐ = Researcher articulated personal beliefs and relation to the conceptualization of the study.
☐ = Research articulated personal beliefs with vague link to conceptualization.
☐ = Research articulated personal beliefs with no link.
☐ = No discussion of researcher’s beliefs related to study’s conceptualization.

A4. Research question guides selection of qualitative, quantitative or mixed methodology of study. This criterion refers to whether a clear link has been established between the research question and methods. In addition, the research methods must be appropriate to the research questions.
☐ = Clear links established between research question and methods, and methods are appropriate to the research question.
☐ = Vague or no links established between research question and methods, but methods are appropriate to the research question.
☐ = Links established between research question and methods, but methods are inappropriate to the research question.
☐ = No links are established and methods are inappropriate to research question.

A5. Participatory Nature of the Research. This criterion refers to the extent to which “stakeholders” were consulted or involved during the research process. Stakeholders refer to those individuals/groups with vested interests or resources relevant to the target problem/goal and to the ecology of the target population; for example, students, teachers, parents, community members, school administrators and support staff, community agencies. “Participation” refers to the level of involvement of participants and other stakeholders in the range of research process activities and decisions, including the development of research questions and methods; data collection and analysis; intervention program design, implementation, and evaluation; and interpretation of data (Nastasi et al., 2004). The purpose of involvement is to ensure that the intervention is culturally specific and socially valid (i.e., incorporates the norms, values, beliefs, practices and behavior relevant to the cultural experiences and social-cultural context of the target population; Nastasi et al., 2004). Participation in each aspect of the research process—generating research questions, formulating methods, implementing the intervention, and interpreting data—is described according to the following criteria.

A5.1. Participatory (Research Questions). Involvement of participants in developing research questions. This criterion addresses the question: Were relevant target group members consulted in the formulation of the research questions?
☐ = Full involvement of target group members (recipients of intervention) in formulating the research questions.
☐ = Involvement of individual representatives of the target group in formulating the research questions.
☐ = Feedback provided by target group members in formulating the research questions.
☐ = No evidence of target group members involvement with formulating research questions.

A5.2. Participatory (Methodology). Involvement of participants in formulating research methods. This criterion addresses the question: Were relevant target group members involved in the formulation of the research methods?
☐ = Full involvement of target group members (recipients of intervention) in formulating the research methods.
☐ = Involvement of individual representatives of the target group in formulating the research methods.
☐ = Feedback provided by target group members in formulating the research methods.
☐ = No evidence of target group members involvement with either formulating research methods.

A5.3. Participatory (Implementation). Involvement of participants in implementation of intervention. This criterion addresses the question: Were relevant target group members involved in the development and delivery of intervention?
☐ = Full involvement of target group members (recipients of intervention) in formulating and implementing the intervention.
= Partial involvement of target group members in intervention—in either formulating or implementing.
= Involvement of individual representatives of the target group in formulating and/or implementing the intervention.
= No evidence of target group members involvement with formulating or implementing the intervention.

A5.4. Participatory (Negotiated Interpretation). Involvement of participants in interpretation of findings. This criterion addresses the question: Were relevant target group members involved in the interpretation of the research findings?

= Full involvement of target group members (recipients of intervention) in interpreting research findings. Interpretation provides authentic representation of participant voices.
= Involvement of individual representatives of the target group in interpreting research findings.
= Feedback provided by target group members about interpretation of results. Representatives are given the opportunity to react to researchers’ interpretations.
= No evidence of target group members involvement with interpreting results.

B. General Design Characteristics

Traditional multiple-participant or group-based studies will be classified according to the type of research design, based on Kirk’s (1995) terminology. If authors provide sufficient information, the categories will be further subdivided as follows:

B1. Random-assignment designs. To receive a code for random-assignment design, authors must specify that random assignment to treatment conditions occurred. If such a statement is not made, then nonrandom assignment is assumed.

B1.1 Completely randomized design, in which at least one type of comparison group is used (e.g., nonintervention, attention control, alternate intervention). Participants are randomly assigned to intervention and comparison conditions.

B1.2 Randomized block design (between-participants variation), in which at least one type of comparison group is used. Participants are initially blocked, or matched, on the basis of some pre-study variable(s) and then randomly assigned to intervention and comparison conditions. For example, a study examining the effects of a classroom level intervention might use this type of design because it is difficult to randomly assign students to a classroom, or classrooms to a school. Instead, the investigators can “match” classrooms based on characteristics such as teacher to student ratio, student reading achievement, student attendance rates, and number of students displaying frequent disruptive behavior. Once two “similar” classrooms are found, or matched, they can be randomly assigned to an intervention or comparison group. This matching process continues until the investigators reach the desired sample size.

B1.3 Randomized block design (within-participants variation), in which at least one type of comparison group is used. Participants are administered both intervention and comparison conditions in different random orders, in either a single-factor layout (i.e., intervention vs. comparison conditions only) or in the context of a multiple-factor design (where the other factors consist of either researcher-manipulated or participant factors). If one or more between-participant factors are combined with one or more within-participant factors, the design is technically referred to as a “split-plot” design, but for present purposes the B1.3 category is sufficient.

B1.4 Randomized hierarchical design. In these designs, treatments are nested instead of being crossed.
B2. Nonrandom-assignment designs. These designs are identical to the random assignment designs described above, except there is no random assignment.

B2.1 Nonrandomized design, in which at least one type of comparison group is used. Participants are nonrandomly assigned to intervention and comparison conditions.

B2.2 Nonrandomized block design (between-participants variation), in which at least one type of comparison group is used. Participants are initially blocked, or matched, on the basis of some pre-study variable and then nonrandomly assigned to intervention and control conditions.

B2.3 Nonrandomized block design (within-participants variation), in which at least one type of comparison group is used. Participants are administered both intervention and comparison conditions in different nonrandom orders, in either a single-factor layout (i.e., intervention vs. comparison conditions only) or in the context of a multiple-factor design (where the other factors consist of either researcher-manipulated or participant factors). If one or more between-participant factors are combined with one or more within-participant factors, the design is technically referred to as a "split-plot" design, but for present purposes the B2.3 category is sufficient.

B2.4 Nonrandomized hierarchical designs. This design is the same as the hierarchical design noted above but random assigned is not used.

B2.5 Optional coding of Quasi-experimental designs (see Appendix E)

B3. Overall confidence of judgment on how participants were assigned.

B3.1 Very low (little basis)
B3.2 Low (guess)
B3.3 Moderate (weak inference)
B3.4 High (strong inference)
B3.5 Very high (explicitly stated)
B3.6 N/A
B3.7 Unknown/unable to code

C. Data Analysis (C1 through C6 are coded for each statistical test that is used in a study).

C1. Appropriate unit of analysis. The unit of analysis corresponds to the unit of intervention (randomization and treatment administration) so that the assumption of independence of errors was not violated and Type 1 error probability was adequately controlled. For example, if a school-wide violence prevention program was being investigated, then whole schools would be the units of analysis—whole schools would be considered the “participant,” and the outcome data would be collected and analyzed at the whole school level or by a hierarchical model that separates school and student level effects. By contrast, individual therapy is administered to individual children; therefore, the unit of analysis would be the child (e.g., the child is the “participant”).

C2. Familywise error rate controlled. There are many ways to control Type I error. However, the purpose of this code is to determine whether some procedure was used to control the familywise or experimentwise Type I error probability when multiple statistical comparisons were made (e.g., use of a MANOVA procedure when multiple related outcomes are evaluated, use of a controlled multiple-comparison procedure used when more than one follow-up comparison is made).

C3. Sufficiently large N. The number of participants (after attrition) in each condition is sufficient to yield enough statistical power for detecting effects of interest or importance. Table 1 provides a guide for the N required for a
small, medium, and large effect size, given the alpha level. When using Table 1, select a statistical test and an alpha level for the analyses in question. This table reflects that larger Ns are needed to detect smaller effects.

C4. **Total size of sample** (start of the study).

C5. **Intervention group sample size.**

C6. **Control group sample size.**
Table 1

*N for Small, Medium and Large ES at Power = .80 for α = .01, .05, and .10*

<table>
<thead>
<tr>
<th>Test</th>
<th>Sm</th>
<th>Med</th>
<th>Lg</th>
<th>Sm</th>
<th>Med</th>
<th>Lg</th>
<th>Sm</th>
<th>Med</th>
<th>Lg</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Mean dif</td>
<td>586</td>
<td>95</td>
<td>38</td>
<td>393</td>
<td>64</td>
<td>26</td>
<td>310</td>
<td>50</td>
<td>20</td>
</tr>
<tr>
<td>2. Sig r</td>
<td>1163</td>
<td>125</td>
<td>41</td>
<td>783</td>
<td>85</td>
<td>28</td>
<td>617</td>
<td>68</td>
<td>22</td>
</tr>
<tr>
<td>3. r dif</td>
<td>2339</td>
<td>263</td>
<td>96</td>
<td>1573</td>
<td>177</td>
<td>66</td>
<td>1240</td>
<td>140</td>
<td>52</td>
</tr>
<tr>
<td>4. P = .5</td>
<td>1165</td>
<td>127</td>
<td>44</td>
<td>783</td>
<td>85</td>
<td>30</td>
<td>616</td>
<td>67</td>
<td>23</td>
</tr>
<tr>
<td>5. P dif</td>
<td>584</td>
<td>93</td>
<td>38</td>
<td>392</td>
<td>64</td>
<td>25</td>
<td>310</td>
<td>50</td>
<td>20</td>
</tr>
<tr>
<td>6. $\chi^2$</td>
<td>1168</td>
<td>130</td>
<td>38</td>
<td>785</td>
<td>87</td>
<td>26</td>
<td>618</td>
<td>69</td>
<td>25</td>
</tr>
<tr>
<td>7. ANOVA</td>
<td>1388</td>
<td>154</td>
<td>56</td>
<td>964</td>
<td>107</td>
<td>39</td>
<td>771</td>
<td>86</td>
<td>31</td>
</tr>
<tr>
<td>8. Mult R</td>
<td>1546</td>
<td>172</td>
<td>62</td>
<td>1090</td>
<td>121</td>
<td>44</td>
<td>880</td>
<td>98</td>
<td>35</td>
</tr>
<tr>
<td>9. $\chi^2$</td>
<td>1675</td>
<td>186</td>
<td>67</td>
<td>1194</td>
<td>133</td>
<td>48</td>
<td>968</td>
<td>108</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>1787</td>
<td>199</td>
<td>71</td>
<td>1293</td>
<td>143</td>
<td>51</td>
<td>1045</td>
<td>116</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>1887</td>
<td>210</td>
<td>75</td>
<td>1362</td>
<td>151</td>
<td>54</td>
<td>1113</td>
<td>124</td>
<td>45</td>
</tr>
<tr>
<td>10. ANOVA</td>
<td>586</td>
<td>95</td>
<td>38</td>
<td>393</td>
<td>64</td>
<td>26</td>
<td>310</td>
<td>50</td>
<td>20</td>
</tr>
<tr>
<td>11. Mult R</td>
<td>464</td>
<td>76</td>
<td>30</td>
<td>322</td>
<td>52</td>
<td>21</td>
<td>258</td>
<td>41</td>
<td>17</td>
</tr>
<tr>
<td>12. ANOVA</td>
<td>388</td>
<td>63</td>
<td>25</td>
<td>274</td>
<td>45</td>
<td>18</td>
<td>221</td>
<td>36</td>
<td>15</td>
</tr>
<tr>
<td>14. ANOVA</td>
<td>299</td>
<td>49</td>
<td>20</td>
<td>215</td>
<td>35</td>
<td>14</td>
<td>174</td>
<td>28</td>
<td>12</td>
</tr>
<tr>
<td>15. Mult R</td>
<td>271</td>
<td>44</td>
<td>18</td>
<td>195</td>
<td>32</td>
<td>13</td>
<td>159</td>
<td>26</td>
<td>11</td>
</tr>
</tbody>
</table>


For studies using qualitative analysis methods, code C7 and C8. It is recommended that these criteria be applied to any coding of qualitative forms of data (e.g., observations, interviews, textual analysis) whether results are represented in qualitative or quantitative form. These criteria are rated for presence/absence (Yes/No).

**C7. Coding.** The use of a systematic and clearly articulated approach to coding of qualitative data that can involve inductive and/or deductive analysis, and meets the following criteria (Nastasi, 1999).

**C7.1** Coding scheme is linked to study’s theoretical-empirical basis.

**C7.2** Procedures for ensuring consistency of coding are used (e.g., dialogue among coders to ensure consistency of meaning and clarify discrepancies; testing the consistency of code application across coders and data).

**C7.3** Progression from abstract concepts to empirical exemplars is clearly articulated.

**C8. Interactive process.** The processes of data collection and analysis interact such that preliminary analytical
results are subjected to further and more focused data collection for the purposes of confirmation, explanation, and validation. The processes of data collection and analysis continue until consistency and depth of understanding are satisfactorily established. This criterion is rated for presence/absence; in addition, the processes used by the researcher are described.

C9. Rival Interpretations In presenting the conceptual basis of the study, the researcher evaluates rival cultural, methodological, statistical, or theoretical explanations and hypotheses. This criterion refers to the extent to which alternative explanations and interpretations of the study’s conceptual foundations and findings are considered by the researchers. This process can be facilitated through interaction with professional colleagues who are not involved in the study (peer review) and/or through participation of the stakeholders throughout the research process. The researcher makes this process explicit by discussing the process and outcome for considering rival interpretations, including how the process was conducted, the alternative perspectives that were considered, and how final decisions were reached.

- 3 = Investigated multiple rival hypotheses (e.g., cultural, methodological, statistical, and/or theoretical explanations) and provided empirical evidence.
- 2 = Investigated one rival hypothesis (e.g., cultural, methodological, statistical, and/or theoretical explanations) and provided empirical evidence.
- 1 = Considered rival explanations without empirical evidence.
- 0 = No evidence that rival hypotheses were investigated or considered (examined data only from one perspective).

D. Type of Program

The prevention or intervention program should be classified according to the Institute of Medicine’s (1994) classification system. Universal prevention programs are designed to forestall the development of mental health problems and to promote competence. Often, in universal programs, it is theorized that everyone would benefit from the program. Selective interventions target individuals or subgroups of a population whose risk of developing a problem is significantly higher than average, or individuals who do not respond to universal prevention programs. Indicated interventions target minimal or detectable signs that foreshadow the development of a problem (IOM, 1994; Shinn, Walker & Stoner, 2002). Intervention or treatments, on the other hand, are provided immediately to treat a serious disorder or problem (Institute of Medicine, 1994).

- D1. Universal prevention program
- D2. Selective prevention program
- D3. Indicated prevention program
- D4. Intervention/Treatment
- D5. Unknown

E. Stage of Program

Prevention or intervention programs will also be classified according to the developmental stage of the program. Rossi and Freeman (1995) classify programs as model/demonstration, early stage, or established/institutionalized programs. Model/demonstration programs are those programs being implemented for the first time, and are usually small in scale and evaluated using the most rigorous methods possible. Fidelity is a major research question at this stage. Evaluations of intervention and prevention programs occurring in the early stages are likely to use mixed methods (e.g., a combination of qualitative and quantitative criteria). Finally, evaluations of established programs are less likely to investigate fidelity of implementation and are more likely to use nonrandomized designs and alternative-intervention comparison groups.

- E1. Model/demonstration programs
E2. Early stage programs
E3. Established/institutionalized programs
E4. Unknown

F. Concurrent or Historical Intervention Exposure

F1. Current Exposure. Participant(s) are exposed to another intervention currently (Specify if information is available).
F2. Prior Exposure. Participant(s) were previously exposed to other interventions (Specify if information is available).
F3. Unknown. No information is available regarding concurrent or historical intervention exposure.
II. Key Features for Coding Studies and Rating Level of Evidence/Support

The second type of coding includes twelve key features of an intervention study that are rated on a four-point scale to specify the level of evidence (i.e., 3=strong evidence/support, 2= promising evidence/support, 1= marginal or weak evidence/support, 0= no evidence/support).

A. Research Methodology

Research methodology pertains to the larger conceptualization of the research used to establish the effectiveness of an intervention and specific issues regarding the methods used to document and report intervention effects. This content includes topics associated with the research design, sampling, selection of measures, descriptions of participants and researchers, and qualities of the measures to establish the effects of an intervention. The conceptualizations of the researchers determine the research questions posed, definitions of constructs, methods of data collection, descriptions of participants, and relationships between researchers and participants, and measurements used. Studies will be evaluated with regard to the qualities of the conceptualizations and measures used to establish the effects of an intervention. Coders will rate the extent to which there is evidence for each of the characteristics or features within the research methodology section. Coding rubrics are delineated for items A1 to A5, following this general progression:

To be rated a 3 for strong evidence, studies must provide evidence for all elements of the criterion specified in this item. Refer to the Coding Protocol for specific elements of this criterion that must be included.

To receive a rating of 2 for promising evidence, evidence must be provided for most of the elements of the criterion specified in this item.

A rating of 1 for weak evidence would require some elements, but not comprehensive or inclusive of most.

A rating of 0 indicates that no evidence of the criterion is available or that the evidence provided is seriously insufficient in the required characteristics or features.

A1. Characteristics of the data collector(s). Studies will be evaluated regarding the level of similarity between data collectors and target participants and the cultural competence of the data collector with respect to the specific target populations.

3 = Provided evidence for high level of similarity to target participants that goes beyond proxy variables (e.g., rapport, trust, communication, ethnic identity) and cultural competence documented.
2 = Provided evidence for some level of similarity to target participants that goes beyond proxy variables (e.g., rapport, trust, communication, ethnic identity) and cultural competence documented.
1 = Provided no evidence for similarity on these variables, but cultural competence is documented.
0 = Provided no evidence for similarity to target participants and/or cultural competence.

A2. Characteristics of participants. Reviewers will evaluate the clarity of descriptions of what culture means for specific groups and the measurement of individuals’ psychological experience or interpretation of cultural membership. For example, to what extent does the study define how culture is conceptualized by participants and how participants interpret their cultural membership? Are psychological experiences, such as acculturation, perceptions of power, oppression operationalized and measured among participants?

3 = Clear definition of what culture means for groups and measurement of individuals’ psychological experience/interpretation of cultural memberships (e.g., acculturation, perceptions of power, oppression).
2 = Clear definition of what culture means for groups without measurement of individuals’ psychological experience/interpretation of cultural memberships individuals’ interpretations (Culture is conceptualized but not operationalized).
1 = Equates culture with demographic labels.
0 = No attempt to describe, define and/or measure culture.
A3. Sample appropriate to research methods. Reviewers will rate the extent that research methods guide sampling procedures (establishing clear links between research method and sampling), and sampling approaches are appropriate to the research methods.

3 = Clear links established between research methods and sampling, and sampling is appropriate to the research methods.
2 = Vague or no links established between research methods and sampling, but sampling is appropriate to the research methods.
1 = Links established between research method and sampling, but sampling is inappropriate to the research method.
0 = No links are established and sampling is inappropriate to research methods.

A4. Operationalization. Specifying the link between key abstract constructs (variables) and data collection methods (operations). Evaluate the clarity of the relationship between key abstract constructs or variables and the data collection methods or operations and the degree that they are operationalized.

3 = Clear links established between constructs and methods, and all key constructs are clearly operationalized.
2 = Some, but not all, key constructs are clearly operationalized.
1 = Vague reference to link between constructs and methods.
0 = No evidence that key constructs are operationalized.

A5. Integration of data from multiple sources, methods, and investigators. Extent to which studies used multiples of all three: sources, methods and investigators.

3 = Used multiple sources, methods, and investigators.
2 = Used two of the following: multiple sources, multiple methods, multiple investigators.
1 = Used one of the following: multiple sources, multiple methods, multiple investigators.
0 = No evidence of multiple sources, methods, or investigators.

B. Measurement

Studies will be evaluated with regard to qualities of the measures used to establish the effects of an intervention. In cases where measures are not well known or not described in enough detail to be evaluated, the coder should contact the author(s) for further information.

To receive a rating of 3 for strong evidence, studies must use measures that produce reliable scores for the majority of primary outcomes under investigation, and for the current population under study. Wallen and Fraenkel (2001) reported for research purposes, reliability should be at least .70, and preferably higher. Since .70 is the rule of thumb for a minimum, a rating of 3 for strong evidence would require that, instruments produce a reliability coefficient of .85 or higher. The reliability information should be provided including the type of reliability statistic used. In addition, to receive a rating of 3, data should have been collected using multiple methods, and collected from multiple sources, when appropriate. Finally, the investigators must have presented a case for the measures used to assess primary outcomes. In the event that multiple primary outcome measures are used, the above criteria must be met for all primary outcome measures.

To receive a rating of 2 for promising evidence, studies must use measures that produce reliable scores for the primary outcomes under investigation, and for the current population under study (i.e., reliability coefficient of at least .70). In addition, data should have been collected either (1) using multiple methods and/or (2) from multiple sources, when appropriate. A case for validity does not need to be presented. In the event that multiple primary outcome measures are used, the above criteria must be met for at least 75% of the primary outcome measures.

To receive a rating of 1 for weak evidence, studies should use measures that produce somewhat reliable scores for the primary outcomes under investigation, and for the current population under study (i.e., reliability coefficient of at least
would get a high score. The goal is to researcher who is deeply involved over a long period of time so that behaviors and thoughts of participants are represented observation. For example, a researcher who comes in one time and collects data would get a lower score whereas a researcher who is deeply involved over a long period of time so that behaviors and thoughts of participants are represented would get a high score. The goal is to attain validity and reliability through trustworthiness, according to Lincoln and Guba (1985).

A rating of 0 indicates that the measures did not produce reliable scores or produced scores with low reliability (<.50), AND/OR data were not collected using multiple methods, AND/OR data were not collected from multiple sources. For example, if the outcome measure is academic achievement and the student's reported achievement is used as the only outcome measure, this would be scored as a 0 (because it is not reliable and only one source was used).

B.1. Use of outcome measures that produce reliable scores. Outcome measures used were supported by evidence of such psychometric characteristics as inter-rater reliability, test-retest reliability, internal consistency, for the population under study. Therefore, when possible, reliability estimates should be reported for the measures used with the samples in the study. This information must either be reported or referenced in the article to receive this code. See Table 2 for information on ways in which reliability can be assessed. Sometimes, investigators combine several measures by using factor analytic procedures to create a single score for a particular construct. This procedure is generally considered a more robust procedure, although the internal consistency of the composite score may not be as strong. In these cases, coders should consider the available reliability and validity evidence and use discretion in their judgment. For example, if researchers combine three scores (α ≥ .85) using confirmatory factor analysis to create a composite score, such that all three scores where derived using multiple methods and from multiple sources, and such that all three scores loaded significantly on one factor, then coders could consider this strong evidence of sound measurement procedures.

Caveat on norm-referenced measures. Often times, researchers use published, norm-referenced measures to assess change. While these instruments are generally considered “reliable,” it is important to note that some norm-referenced assessments may not produce reliable scores for certain populations. For example, some studies may use well-known, norm-referenced measures on populations not well represented in the norming process. Even if the population in question was represented in the norming process, it does not necessarily mean the instrument produced reliable scores for that group. In addition, researchers will sometimes use less reliable subscales of published norm-referenced instruments as key outcome measures. Ideally, researchers should have calculated a reliability coefficient for all instruments used to measure change to ensure that the measure produced reliable scores for the population under study, even when well-known and widely used measures were used. When authors do not provide these data, coders would consider the absence of that criterion when making their reliability rating.

B.2. Multi-method. Multiple (i.e., at least two) assessment methods or approaches were used (e.g., observational, self-reports, teacher ratings). This rating is not always applicable. For example, directly observable incidence rates may not require a multi-method assessment.

B.3. Multi-source. Measures were obtained from multiple (i.e., at least two) sources (e.g., teachers, parents, self). This rating is not always applicable. Directly observable incidence rates would not necessarily require a multi-source assessment.

B.4. Extent of Engagement. Level of engagement affects both (a) the depth and reliability of representations, and (b) the scope and depth of understandings and relationships within the study. High levels of engagement assure deep and accurate representations and understandings, thereby affecting understanding and reliability in the study of intervention outcomes. Prolonged engagement means that the data are collected over a substantial period of time to ensure accuracy of representations. Persistent observation means that the data collection is progressively focused to ensure thorough understanding of consistency and variation likely to occur in intervention process and outcomes. The goal is for researchers to collect data in a manner that guarantees sufficient scope and depth through prolonged engagement and persistent observation. For example, a researcher who comes in one time and collects data would get a lower score whereas a researcher who is deeply involved over a long period of time so that behaviors and thoughts of participants are represented would get a high score. The goal is to attain validity and reliability through trustworthiness, according to Lincoln and Guba (1985).

3 = Provided evidence for high level of engagement to ensure deep and accurate representation.
2 = Provided evidence for some level of engagement to ensure deep and accurate representation.
1 = Provided evidence of minimal level of engagement to ensure deep and accurate representation.
0 = Provided no evidence for level of engagement to ensure deep and accurate representation.

B.5. Validity of measures reported. Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Validation involves gathering evidence to provide a sound scientific basis for the interpretation of scores. It is important that authors note the way in which validity of their primary outcome measures was assessed. This criterion also requires attention to validity relevant to the target population. Thus, to receive the maximum score, the researchers must present evidence that measures have been validated with the target group. Table 2 provides information on different types of validity that can be assessed, as well as some important cultural considerations in assessing validity. Observable incidence and/or occurrence rates, such as school attendance rates and homework completion rates, and well-known standardized, norm-referenced assessments will be considered valid measures. However, coders should consider the cultural considerations listed in Table 2 and use discretion in coding measure validity.

3 = Yes, validated with specific target group.
2 = In part, validated for general population only.
1 = No.
0 = Unknown/unable to code.

B.6. Cultural Appropriateness of the Measures. Studies will be evaluated on the extent to which evidence is provided to demonstrate the cultural appropriateness of the measures for the target group and the methods by which this determination was made. In rating this item, consider the following dimensions: meaning, language, dialect, and response format used in the methods of data collection and responses. For example evaluate evidence that participants provided data in a way that is culturally familiar and in their specific language and dialect. Record the extent to which there is direct empirical evidence that the measures are in a format that is culturally appropriate for this specific target group versus evidence from related or other groups. A maximum rating for cultural appropriateness indicates that measures were developed through formative research with the target group.

3 = Developed measure for use with target group in study on the basis of empirical evidence (conducted formative research and developed measure).
2 = Adapted existing measure for use with target group on the basis of formative research and/or empirical evidence with target group.
1 = Developed or adapted measures for use with target group based on empirical evidence with similar or related populations.
0 = Measure not tailored specifically for target group.

B7. Measures of key outcomes are linked to the conceptual model. The selection of key outcome variables and indicators are linked to the conceptual model (grounded in theory and research). The connections between outcome indicators and variables in the conceptual model are clearly articulated. Outcome indicators reflect different points in the pathways of change (as articulated in the conceptual model) such as independent, mediating, and dependent variables. For example, in the conceptual model, the independent variable (peer norms) is linked to the mediating variable (perceptions of risk) which is linked to the dependent variable—risky sexual behavior. The proposed intervention is directed toward reducing risky behavior by influencing peer norms and risk perception. Indicators for all three variables should be included in order to test proposed pathways of change.

3 = Clear links established between the conceptual model and key outcome indicators.
2 = Some, but not all, key outcomes are clearly linked to conceptual model.
1 = Vague reference to links between key outcomes and conceptual model.
0 = No evidence that key outcomes are linked to conceptual model.
### Table 2

**Methods of Checking Validity and Reliability**

<table>
<thead>
<tr>
<th>Validity Type</th>
<th>Cultural Considerations</th>
<th>Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence based on test content (Content Validity)</td>
<td>Differences in language and meaning</td>
<td>Expert judgment</td>
</tr>
<tr>
<td></td>
<td>Relevance of items for different cultural groups</td>
<td>Translation and back translation procedures</td>
</tr>
<tr>
<td></td>
<td>Culturally defined response styles</td>
<td></td>
</tr>
<tr>
<td>Evidence based on response processes</td>
<td>Individuals are responding to the measure as represented by the construct</td>
<td>Conduct an analysis of individual items</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Question responders about the meaning of their responses</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Document related aspects of performance as related to the construct being measured</td>
</tr>
<tr>
<td></td>
<td>Analyze the relationship among parts of the test and between the test and other constructs/variables</td>
<td>Assess how observers/judges record and evaluate data related to the response processes</td>
</tr>
<tr>
<td>Evidence based on relations to other variables (Criterion Validity)</td>
<td>Accuracy of predictions across different cultural groups</td>
<td>Relate to another measure of the same variable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Relationship between tests scores and relevant criteria are modeled by regression lines and compared across groups.</td>
</tr>
<tr>
<td>Evidence based on internal structure (Construct Validity)</td>
<td>Construct defined in a culturally relevant fashion</td>
<td>Assess evidence on predictions made from theory</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Factor analysis conducted using data collected from population under investigation.</td>
</tr>
<tr>
<td>Evidence based on consequences of testing</td>
<td>Individual benefits will be derived from the intended use of scores (e.g., beneficial interventions)</td>
<td>Assess evidence of differential benefits of measures for specific groups</td>
</tr>
</tbody>
</table>
**Reliability ("consistency")**

<table>
<thead>
<tr>
<th>Type</th>
<th>Content</th>
<th>Time Interval</th>
<th>Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test-retest</td>
<td>Identical</td>
<td>Varies</td>
<td>Give identical instrument twice</td>
</tr>
<tr>
<td>Equivalent forms</td>
<td>Different</td>
<td>None</td>
<td>Give two forms of instrument</td>
</tr>
<tr>
<td>Internal Consistency</td>
<td>Different</td>
<td>None</td>
<td>Divide instrument into halves and score each or use K-R</td>
</tr>
<tr>
<td>Observer agreement</td>
<td>Identical</td>
<td>None</td>
<td>Compare scores obtained by two (or more) observers</td>
</tr>
</tbody>
</table>


**C. Comparison Group**

In judging the merit of an intervention, there must be a comparison group of some type against which to compare the intervention. The intervention in question could have been shown to be superior to a no intervention or wait list control, an attention or placebo control, or an alternate intervention. If the intervention in question was compared to an alternate intervention and to a control condition in one study, it may be shown superior to the control condition but not the alternate intervention. Evidence ratings will consider all available information related to the comparison group findings.

To be rated a **3 for strong evidence**, at least one type of "active" comparison group (e.g., typical intervention, attention placebo, intervention element placebo, alternative intervention, pharmacotherapy) must be used. Initial group equivalency must be established, preferably through random assignment of participants to intervention conditions. There must be evidence that change agents were counterbalanced, as well as the study must meet the criteria for equivalent mortality and low attrition at post, and if applicable, at follow-up.

To receive a rating of **2 for promising evidence**, at least a "no intervention group" type of comparison must have been used (e.g., no intervention, wait list/delayed intervention, minimal contact). In addition, there must be evidence for at least two of the following: (1) counterbalancing of change agents, (2) group equivalence established, or (3) equivalent mortality with low attrition. If equivalent mortality is not demonstrated, an intent-to-intervene analysis must have been conducted, resulting in a finding of no significant group differences between the control and treatment groups.

A rating of **1 for weak evidence** would require a comparison group and at least one of the following: (1) counterbalancing of change agents, (2) group equivalence established, or (3) equivalent mortality with low attrition. If equivalent mortality is not demonstrated, an intent-to-intervene analysis must have been conducted, resulting in a finding of no significant group differences between the control and treatment groups.

A rating of **1** would require a comparison group (e.g., a waiting list); however, no group equivalence procedures were used (according to the methods described next).

A rating of **0** indicates that no efforts were made to ensure group equivalence.
C1. Type of comparison group will be classified as:

C1.1 Typical intervention: control receives a typical intervention in a school setting without special additions that constitutes the intervention of interest.
C1.2 Typical intervention (other): Specify.
C1.3 Attention placebo: control receives attention, discussion, or a weaker form of the intervention.
C1.4 Intervention element placebo: control receives target intervention minus the active ingredient linked to therapeutic change.
C1.5 Alternative intervention: Control receives another intervention (other than typical intervention) or at a lower dosage than the established intervention (e.g., planned variation studies). This is applicable when the alternative intervention is not expected to be as effective as target intervention.
C1.6 Pharmacotherapy.
C1.7 No intervention: no evidence of treatment or attention.
C1.8 Wait list/delayed intervention: contact limited to application, screening, pre or posttest.
C1.9 Minimal contact: includes participation in intake interview, given instructions, but not wait listed.
C1.10 Unable to identify type of comparison group.

C2. Overall confidence rating on judgment of type of comparison group

C2.1 Unable to identify comparison group.
C2.2 Very high (explicitly stated).
C2.3 High (strong inference).
C2.4 Moderate (weak inference).
C2.5 Low (guess).
C2.6 Very low (little basis).

C3. Counterbalancing of change agents. All participants who received intervention from a single therapist/consultant/mediator/instructor or multiple agents were counterbalanced across intervention conditions to avoid agent-by-condition confound. This coding means, for example, that any effects an individual therapist/consultant/teacher/mediator may have produced in the intervention condition also affected participants in the comparison condition. A statistical analysis that controls for “change agent” or “classroom” effect is acceptable.

C4. Group Equivalence Established

C4.1 Random assignment. To be coded for random assignment, participants must have been assigned randomly to intervention and control groups in group-based designs. Assignment may have been completely random, or random with blocking on, for example, level of treated problem, age, social economic level, pretest measures and/or IQ scores, number of suspensions, or other variables deemed important by the study authors. Threats to internal validity that random assignment does not rule out are considered in determining the level of evidence to support random assignment.

C4.2 Post hoc matched set. Participants were paired and matched on a given set of characteristics such as age, gender, ethnicity, ability, etc., and then assigned randomly to conditions.

C4.3 Statistical matching. This procedure is accomplished by adjusting participants’ scores by one or more covariates.

C4.4 Post hoc test for group equivalence. Statistical tests were conducted to demonstrate that, prior to the intervention, the intervention and control group(s) were equivalent. One must be cautious to have enough power with this procedure.

C5. Equivalent mortality with low attrition (less than 20% for post; 30% for follow-up). The number of participants
in each treatment group at the conclusion of the study was approximately the same (i.e., less than 20% attrition at post assessment and less than 30% at follow-up assessment). Nonequivalent mortality rates may be associated with negative treatment characteristics or important differences between those participants who remain in the study and those who withdraw.

Levels of attrition will be reported for all groups/participants in the study, to the extent that the study authors provide this information. Reviewers will also note how attrition was accounted for in the data analysis (e.g., initial versus terminal intervention sample as well as nonintervention dropouts, analysis intent-to-intervene participants or completers only). Reviewers will report on whether intent-to-intervene analyses were carried out, and if so, what the findings were and how they differed from findings of "as-intervened" analyses.

D. Primary/Secondary Outcomes Are Statistically Significant

One important criterion is whether the intervention demonstrates a statistically significant effect on the outcome measures in the intervention group or condition. The Task Force will consider the guidelines presented by Wilkinson and the Task Force on Statistical Inference (1999) of the APA Board of Scientific Affairs. Although statistical significance is a necessary criterion for intervention outcome efficacy, it is not sufficient in that statistically significant effects do not guarantee effects of practical importance. The statistical significance, the effect sizes, and whenever possible, the statistical power, from all primary outcome measures will be assessed in judging the evidence to support a primary outcome. Outcomes will be considered statistically significant if they reach an alpha level of .05 or less. Familywise and experimentwise error rates will be considered.

To be considered a "primary outcome," the outcome measure should represent a valid and appropriate indicator for the type of prevention or intervention. In other words, "primary outcomes" are those that reflect the ultimate goal(s) of the intervention. For example, for interventions that target a behavior-linked problem such as drug use or sexual activity, at least one outcome measure must assess behavioral functioning or adjustment status; an outcome restricted to knowledge assessment would not be sufficient. Similarly, if an intervention targets reading improvement, than the outcome must measure reading performance. For both of these intervention types, outcomes restricted to attitudes would not be sufficient to meet the criteria of a primary outcome. Outcomes significant at the $p < .05$ level will be described in the "Primary/Secondary Outcomes Significant" table. Reviewers will note any nonsignificant and/or negative outcomes associated with the intervention in a separate table (see "Null Findings/Negative Outcomes").

To receive a rating of 3, strong evidence, an appropriate statistical analysis must have been conducted, including appropriate units of analysis, familywise/experimentwise error rate controlled (if applicable), and a sufficiently large $N$. A study must show significant primary outcomes for at least 75% of the total primary outcome measures for each key construct. Please note that "75%" refers to a single primary outcome construct, not a sum of the total primary outcome measures. For example if an increase in reading performance and reduced disruptive behavior were identified as primary outcomes, and there were 2 measures of reading and 4 measures of disruptive behavior, then both measures of reading and 3 of 4 measures of disruptive behavior must demonstrate statistical significance in the desired direction to receive a rating of 3, strong evidence. Measured outcomes must also reflect a moderate effect size (See Tables 3-6). In addition, if primary outcomes are dependent upon related outcomes as suggested by the underlying theory, statistical significance must be demonstrated for these outcomes. If a statistical procedure indicating the magnitude of an intervention effect was not calculated, the coder should attempt to compute it based on available data.

To receive a rating of 2, promising evidence, an appropriate statistical analysis must have been conducted, including appropriate units of analysis, familywise/experimentalwise error rate controlled (if applicable). A study must show significant and primary outcomes for at least 50% to 74% of the total primary outcome measures for each key construct (e.g., at least 2 out of 4 primary outcome measures for a given construct show statistically significant change in the desired direction). In addition, if primary outcomes are dependent upon related outcomes as suggested by the underlying theory, statistical significance must be demonstrated for these outcomes.

To receive a rating of 1, weak evidence, an appropriate statistical analysis must have been conducted, including appropriate unit of analysis, familywise/experimentalwise error rate controlled (if applicable). A study must show significant and primary outcomes for between 25% and 49% of the total primary outcome measures for any key construct (i.e., at least 1 out of 4 outcome measures show statistically significant change in the desired direction for a given primary outcome.
A rating of 0 would indicate that none of the above criteria were met. Significant primary outcomes will be coded according to the following categories:

**D1. Evidence of appropriate statistical analysis for primary outcomes.** Coders will consider whether an appropriate statistical analysis was conducted using the information gathered in Statistical Treatment (Section I, Criterion C). In addition, coders will consider any glaring errors and oversights in the statistical analysis that may not have been captured in the Statistical Treatment section (e.g., an obvious violation of an assumption, such as nonindependence of error components in an Analysis-of-Variance).

- **D1.1 Appropriate unit of analysis.**
- **D1.2 Familywise/experimentwise error rate controlled.**
- **D1.3 Sufficiently large N.**

**D2. Percentage of primary outcomes that is statistically significant.**

- **D2.1 Outcome.** Specify the outcome being tested, noting primary outcomes before secondary outcomes. All outcomes are listed in alphabetical order.

- **D2.2 Primary vs. Secondary.** Specify whether the outcome was primary or secondary.

- **D2.3 For whom the outcome(s) were significant**
  - Child.
  - Teacher.
  - Parent/significant adult.
  - Ecology (e.g., classroom context/peers).

- **D2.4 What changed**
  - Behavior: Specify.
  - Attitude: Specify.
  - Knowledge: Specify.
  - Dynamics: Specify.
  - Other: Specify.

- **D2.5 Source/type of measurement**
  - Self-report: Specify.
  - Parent report: Specify.
  - Teacher report: Specify.
  - Peer ratings/nominations: Specify.
  - Observation: Specify.
  - Test: Specify.
  - Other: Specify.
  - Unknown/information not provided.

- **D2.6 Treatment Information.** Specify which treatment groups are being compared.

- **D2.7 Outcome measure used and reliability.** Specify which measure was used to investigate the outcome, and list the reliability. If subscales were analyzed individually, then treat them as individual measures of a given construct. If measures or scales were combined, then treat the combined score as a single measure of a given construct, and list the reliability accordingly.

- **D2.8 Effect Size.** The effect size will be recorded as provided by authors. If the authors do not provide the
effect size, then the effect size will be calculated by the reviewer(s). (See Tables 3-6 for instructions on how to calculate selected effect sizes and magnitude guidelines.) In addition, the type of data the effect size is based on will be reported, as well as the strength of the statistic (e.g., small, medium and large). Appendix E also provides general resources that will assist reviewers in calculating and interpreting effect sizes.

D2.9 Power. Statistical power considerations should be taken into account.

D3. Evidence of appropriate statistical analysis for secondary outcomes. Empirical evidence and/or theory suggests that these outcomes are necessary for, or are associated with, attaining the primary outcomes. These issues may be especially important to consider for program adoption by practitioners or researchers engaged in replication work. For example, an intervention might be designed to target challenging behavior in the classroom by including a component designed to increase the number of times praise is delivered and improve the specificity/quality of the praise. According to this example, reducing challenging behavior in the classroom is the primary outcome, and increased quality and quantity of teacher praise is the secondary outcome necessary or associated with achieving the primary outcome. Distinguishing between Primary and Secondary Outcomes can become difficult. To simplify the distinction, coders should rely on the explanation provided by the authors regarding the mechanism through which change in the primary outcomes will occur. If the authors do not provide such an explanation, secondary outcomes will not be coded. Secondary Outcomes will be coded along the same dimensions as Primary Outcomes.

D4. Percentage of secondary outcomes that is statistically significant

D5. Overall Summary of Questions Investigated

D5.1 Main effect analyses. Reviewers will note whether main effect questions were investigated in the study (e.g., the average effect of participation in the intervention/prevention program on primary and secondary outcomes).

D5.2 Moderator effect analyses. Reviewers will note whether the authors investigated subgroup effects (e.g., did the intervention/prevention program produce stronger effects for certain subgroups). Reviewers will specify the results of those analyses on the coding sheet when applicable.

D5.3. Mediator analyses. Reviewers will note whether the authors investigated factors or pathways that mediate program outcomes. Mediator analyses help explain the link between program activities and program outcomes (see Reynolds, 2000 for additional information on mediator analyses). Reviewers will specify the results of those analyses on the coding sheet when applicable.

D5.4. Unintended Outcomes. Evidence of unintended positive and negative outcomes is provided. In-depth documentation of implementation is expected to yield information about unintended outcomes. In describing unintended outcomes, follow the same guidelines as those for intended outcomes.

D6. Cultural Moderator Variables. Investigated effects of cultural moderating variables on the outcome. Attention to moderator variables includes consideration of the cultural variables that may impact program implementation and outcomes. Moderator variables include the perceptions of the participants regarding the acceptability and cultural validity of the intervention; that is, whether the intervention content and process is consistent with their cultural perspectives and experiences, and whether they find the intervention to be feasible and enjoyable, and beneficial for their daily lives. These considerations, in conjunction with conditions for implementation, are critical for and interpreting the outcomes, and determining the likelihood of sustainability and transferability of interventions.

3 = Investigated effects of cultural variables, with culture operationalized with reference to acculturation, perceptions of power, oppression.
2 = Investigated effects of cultural variables, with culture operationalized as demographic variables (e.g., race, ethnicity, language).
1 = Interpreted results in terms of culture without formal investigation, but with reference to other research with
similar or related cultural groups.
0 = No evidence that moderator effects were investigated and no reference to influence of culture in interpretation of findings.

Table 3

**ES Indexes and Their Values for Small, Medium, and Large Effects**

<table>
<thead>
<tr>
<th>Test</th>
<th>ES index</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. (m_A) vs. (m_B) for independent means</td>
<td>(d = \frac{m_A - m_B}{\sigma})</td>
<td>Small 0.20 Medium 0.50 Large 0.80</td>
</tr>
<tr>
<td>2. Significance of product moment (r)</td>
<td>(q = zA - zB) where (z = \text{Fisher's } z)</td>
<td>(R) Small 0.10 Medium 0.30 Large 0.50</td>
</tr>
<tr>
<td>3. (r_A) vs (r_B) for independent (r)s</td>
<td>(g = P - .50)</td>
<td>(g) Small 0.05 Medium 0.15 Large 0.25</td>
</tr>
<tr>
<td>4. (P = .5) and the sign test</td>
<td>(h = \phi_A - \phi_B) where (\phi = \text{arcsine transformation})</td>
<td>(h) Small 0.20 Medium 0.50 Large 0.80</td>
</tr>
<tr>
<td>5. (P_A) vs. (P_B) for independent proportions</td>
<td>(w = \sum_{i=1}^{k} \frac{\left(\frac{P_{ii} - P_{0i}}{P_{0i}}\right)^2}{P_{0i}})</td>
<td>(w) Small 0.10 Medium 0.30 Large 0.50</td>
</tr>
<tr>
<td>6. Chi-square for goodness of fit and contingency</td>
<td>(f = \frac{\sigma_m}{\sigma})</td>
<td>(f) Small 0.02 Medium 0.15 Large 0.35</td>
</tr>
<tr>
<td>7. One-way analysis of variance</td>
<td>(f^2 = \frac{R^2}{1 - R^2})</td>
<td>(f^2) Small 0.02 Medium 0.15 Large 0.35</td>
</tr>
</tbody>
</table>


Table 4

**Formulas to Compute Various Measures of Effect Size**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Effect Size</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>(F)</td>
<td>(d) index</td>
<td>(2t / \sqrt{df})</td>
</tr>
<tr>
<td>(F)</td>
<td>(f) index</td>
<td>(\sqrt{df_{effect}(F) / df_{error}})</td>
</tr>
<tr>
<td></td>
<td>Eta squared (\eta^2)</td>
<td>(df_{effect}(F) / (df_{effect}(F) + df_{error}))</td>
</tr>
<tr>
<td>(\chi^2)</td>
<td>(\varphi) index</td>
<td>(\sqrt{\chi^2 / N})</td>
</tr>
<tr>
<td></td>
<td>Cramer’s V</td>
<td>(\sqrt{\chi^2 / N {Min(r - 1, c - 1)}})</td>
</tr>
<tr>
<td>(R)</td>
<td>(r) index</td>
<td>Correlation coefficient</td>
</tr>
<tr>
<td></td>
<td>(t) or (r)</td>
<td>(r = t / \sqrt{t^2 + df})</td>
</tr>
</tbody>
</table>

Table 5

**Ranges for Effect Size Indexes**

<table>
<thead>
<tr>
<th>Effect Size</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>d index</td>
<td>.20</td>
<td>.50</td>
<td>.80</td>
</tr>
<tr>
<td>f index</td>
<td>.10</td>
<td>.25</td>
<td>.40</td>
</tr>
<tr>
<td>r index</td>
<td>.10</td>
<td>.30</td>
<td>.50</td>
</tr>
<tr>
<td>φ index</td>
<td>.10</td>
<td>.30</td>
<td>.50</td>
</tr>
</tbody>
</table>


Table 6

**Formulas to Convert Various Effect Size Estimates into d Index Values**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>( d = 2t / \sqrt{df_{error}} )</td>
</tr>
<tr>
<td>F(^a)</td>
<td>( d = 2F / df_{error} )</td>
</tr>
<tr>
<td>r</td>
<td>( d = 2r / \sqrt{1 - r^2} )</td>
</tr>
<tr>
<td>χ(^b)</td>
<td>( d = 2\sqrt{\chi^2 / N / \sqrt{1 - \chi^2 / N}} )</td>
</tr>
</tbody>
</table>

*Note.* For additional conversion formulas see Cohen (1988) \(^a\)F to d index conversion for single degree of freedom F-test. \(^b\)χ\(^2\) to d index conversion for 2 x 2 contingency table. From “Practical Significance In Early Intervention Research: From Affect To Empirical Effect,” by K. J. Ottenbacher, 1992, *Journal of Early Intervention, 16*, 184. Copyright 1992 by the Division for Early Childhood, the Council for Exceptional Children. Adapted with permission of the author.
E. Cultural Significance

This criterion refers to the cultural relevance or significance (i.e., cultural validity) of key outcome measures. *Culture* is defined as the shared norms, values, beliefs, practices, and behaviors relevant to the target population; and reflects cultural experiences that extend beyond the traditional categories of race, ethnicity, and language to encompass the shared experiences of any group (e.g., culture of classroom, school, neighborhood, family, peer group). The selection of key outcome variables and indicators is based on formative research that documents relevance to ecological/social contexts and the cultural experiences of the target population and the participants. The indicators are meaningful and ecologically valid given the context and experiences of the participants. The indicator has relevance to the real-life experiences of the target population and the participants as supported by research evidence. For example, in research on prevention of sexually transmitted diseases, formative research has established that condom use is a viable and culturally-appropriate alternative for preventing HIV infection in the target population and for the specific participants. Thus, the indicator represents concepts that have legitimacy within the shared social-cultural experiences of the target population and participants. Ratings for this criterion are based on the extent to which the relevance to the target population has been established and the nature of the evidence.

- 3 = Conducted formative research or provided empirical evidence with the target population to show relevance of key outcome variables and indicators to target population.
- 2 = Cited empirical evidence for relevance of key outcome variables and indicators to target population.
- 1 = Cited empirical evidence with similar or related populations for relevance of key outcome variables and indicators to target population.
- 0 = No empirical evidence cited for relevance of outcome variables and indicators to population.

F. Educational/Clinical Significance

Clinical significance is defined as changes in an individual’s behavior that are meaningful and important (Kazdin, 1977). Furthermore, changes in behavior must “resolve or significantly ameliorate the problem for which the client sought treatment” (Kazdin, 1977, p. 427). Clinical significance includes four domains that should be considered when evaluating the quality of evidence related to educational/clinical significance.

- F1. Categorical diagnosis data. Criteria may be provided regarding inclusion into the study, or changes in diagnostic criteria at pre, post, and follow-up.
- F2. Outcomes assessed via continuous variables. This criterion is defined as the percentage of participants who fall inside the range of dysfunction, the percentage of participants who fall in the normal range, or the percentage of participants showing clinical improvement during each phase of the intervention.
- F3. Subjective evaluation. Reviewers should code any information provided regarding the importance of the behavioral change at pre, post, and follow up as evaluated by individuals in direct contact with the participant.
- F4. Social comparison. Reviewers should code any information relating to the participant’s behavior at pre, post, and follow up in comparison to normative data such as typical peers.

To receive a rating of 3, strong evidence, a study must provide evidence in support of the clinical significance for at least 3 of the 4 criteria listed (i.e., Categorical Diagnosis Data, Outcomes Assessed Via Continuous Variables, Subjective Evaluation, or Social Comparison) during either post or follow up phases for the majority of participants.

To receive a rating of 2, promising evidence, a study must provide evidence in support of the clinical significance for at least 2 of the 4 criteria listed during either post or follow up phases for the majority of participants.

To receive a rating of 1, weak/marginal evidence, a study must provide evidence in support of the clinical significance for at least 1 of the 4 criteria for the majority of participants.
A rating of 0 indicates that none of the above criteria were met.

G. External Validity Indicators

G1. Participant selection (sampling). Participants refer to all persons involved in the intervention, including recipients (e.g., students), implementers (e.g., teachers, psychologists), staff trainers (e.g., researchers, interventionists), decision makers (e.g., principal, research team members), and other key stakeholders who may influence the other participants (e.g., parents, community members, superintendent). Participants are described in sufficient detail to permit other researchers and interventionists to determine both the conditions under which the intervention occurred and the generalizability to intended participants. Reviewers will consider any information reported by the authors regarding the sources from which the sample was drawn, and the means by which clients came to be in the study (e.g., screening at an elementary school, university clinic pool of families, newspaper advertisements, community mental health referral).

G1.1 Sampling procedures described.
G1.2. Rationale for sample selection specified
G1.3 Rationale for sample size specified
G1.4 Evidence provided that sample represents the target population
G1.5 Recruitment procedures are congruent with target cultural group. Reviewers will determine whether culturally appropriate ways were used to contact, recruit, inform, and maintain participation of the participants.

G1.6 Inclusion/exclusion criteria specified. Reviewers will consider what participant selection criteria the authors applied to the sample that might limit its representativeness (e.g., whether authors required parent participation, required a two-parent family, or whether they excluded comorbid cases).

G1.7 Inclusion/exclusion criteria similar to school practice. Reviewers will also consider whether inclusion/exclusion criteria were similar to those likely to be used in school practice (code as yes/no). For example, criteria that are similar to school practice might include ruling several physically aggressive students out of a prevention program, and excluding youth from individual psychotherapy who are currently abusing substances. Criteria that are dissimilar from school practice might involve, for example, excluding children who do not have two parents living at home, or eliminating cases with comorbid diagnoses other than the one targeted in intervention.

G1.8 Specified criteria related to concern. The criteria for inclusion into the study should be related to the goal of the intervention. For example, if an intervention is designed to reduce aggressive behavior, then some specified criterion related to the presence of aggressive behavior should be included in the study.

G2. Participant characteristics specified for treatment and control group. Characteristics of the samples used to test interventions will be specified. An objective here is to identify the range of client (e.g., child/family/mediator) characteristics within which an intervention has been shown to have beneficial effects. Each intervention and control sample will be described along the following demographic dimensions:

G2.1 Type of participant. Specify student, parent, teacher, or larger system.
G2.2 Grade/age
G2.3 Gender
G2.4 Ethnicity or multiethnic
G2.5 Ethnic identity
G2.6 Race(s)
G2.7 Acculturation
G2.8 Primary language
G2.9 SES
G2.10 Family structure (e.g., single parent, immigrant status).
G2.11 Locale (e.g., urban, rural, suburban, university-affiliated site).
G2.12 Disability (or other special considerations such as English Language Learners).
G2.13 Functional descriptors (e.g., clinical, sub-clinical, "normal" groups; other academic problems present, such as at-risk factors, low reading achievement, etc.).

G3. Details are provided regarding demographic variables (age, gender, and ethnicity-race) and any other variables that have been identified as:

G3.1. Having differential relevance for intended outcomes.

G3.2. Being relevant to inclusion or exclusion (e.g., level of education, prior experience).

G5. Transferability of intervention. Level of detail provided to permit decisions about transferability (including decisions about similarity of context, participants, intervener, and available resources; key characteristics of the intervention are transportable). Decisions about transferability (external validity) of the intervention are dependent on “rich description” of the study’s conditions. Rich description involves in-depth detailing of the procedures and conditions surrounding implementation—including delivery of the intervention, training of facilitators, monitoring implementation, and making necessary adaptations. The documentation of the procedures and conditions surrounding implementation and adaptation must be of sufficient detail to permit other researchers and practitioners to transfer the intervention to other contexts and participants, and to determine the applicability of fit to intended contexts and participants. Thick description should include descriptions about how decisions regarding adaptation to local conditions were made during the course of the intervention study. Decisions about transportability or “generalization” are ultimately within the purview of the consumer; however, the researcher’s responsibility is to provide sufficient detail to facilitate the consumer’s decision making about applicability and provide a process for adapting the intervention to local conditions.

3 = Complete and detailed description of the context within which the intervention occurs.
2 = Detailed description of some but not all contextual components.
1 = Provides overview of contextual components but lack details.
0 = No description of context.

G6. Participant Perception of Benefits of the Intervention (Receptivity/acceptance by target participant population). Investigated perceptions of benefit for all participant groups and reported participants benefiting overall from the intervention. Please note any consumer satisfaction information, such as reactions from children, parents, teachers, or other program participants as well as reactions to and perceptions of what was gained as a result of the program. Reviewers should also provide a global rating as to whether the participants reported benefiting or not benefiting from the intervention.

3 = Provided evidence of perceived benefits from the intervention for all participant groups.
2 = Provided evidence of perceived benefits from the intervention for some participant groups.
1 = Provided evidence that participants did not perceive benefits from the intervention.
0 = Did not investigate participants’ perceptions of benefits.

H. Durability/Generalization of Intervention and Outcomes

Durability or generalization is generally defined as the length of time over which change was maintained, that is, the extent to which outcomes generalize over time. It involves evidence of continued improvement and supporting effects after a follow-up interval. From a practical perspective, durability relates to whether change was maintained once support from the research team (or intervention agent) was withdrawn. Such ratings are based on information authors provide on the number and timing of outcome assessments. Generally, the longer an effect of the intervention lasts, the more benefits it should generate for individuals or groups participating in the program.
**Feature H1** (Follow up Assessment) addresses the extent to which follow-up assessment was built into the design. The rating is based on information the authors provide on the timing of outcome assessments, the number of participants included in the follow up assessment, in addition to the consistency of assessment method used.

For the purposes of examining durability, we have used an expanded definition that addresses maintenance of change over time, settings, and persons. In addition, we have included criteria that address whether the authors provide evidence of durability/generalization and whether they have incorporated strategies for ensuring durability/generalization across time, setting, and persons. **Features H2-H4** addresses these criteria.

Follow-up activities help to determine not only the extent to which intervention outcomes are sustained, but also the extent to which intervention efforts are maintained (continued beyond the designated intervention/research project period; i.e., program sustainability) by the local stakeholders and the extent to which the organization or system invests sufficient resources or has the capacity to sustain efforts (i.e., institutionalization). We have included criteria that address both program sustainability (H2.2) and institutionalization (H3.3).

Coders will rate durability/generalization features for presence/absence of evidence (yes/no) and provide descriptive information. In addition, coders are asked to provide an overall rating on follow-up assessment (H1) using a 4-point scale.

**H1. Follow Up Assessment.**

Reviewers will code the extent to which follow up assessment was built into the study design. The rating is based on information the authors provide on the timing of outcome assessments, the number of participants included in the follow up assessment, in addition to the consistency of assessment method used.

To receive a rating of **3**, **strong evidence**, the study must have conducted follow up assessments over multiple intervals (e.g., 6 months, 1 year), with all participants that were included in the original sample, using similar measures used to analyze data from primary or secondary outcomes.

To receive a rating of **2**, **promising evidence**, the study must have conducted follow up assessments at least once (e.g., 6 months), with the majority of participants that were included in the original sample, using similar measures used to analyze data from primary or secondary outcomes.

To receive a rating of **1**, **weak evidence** at least once (e.g., 6 months), with some participants from the original sample.

A rating of **0**, no evidence, would indicate that no follow up assessment was built into the study.

**H2 Durability/Generalization over time.**

**H2.1 Persistence (sustainability) of outcomes.** Evidence is provided regarding the extent to which outcomes persist following the termination of the intervention.

**H2.2 Procedures for ensuring sustainability.** The conditions under which sustainability has been achieved; that is, documentation of efforts (or lack thereof) to ensure maintenance of outcomes (e.g., through booster sessions).

**H3 Durability/Generalization across settings.**

**H3.1 Application outside of the intervention context.** Evidence is provided regarding the extent to which outcomes are manifested in contexts that are different from the intervention context. For example, evidence of the capacity for the participants to make use of newly acquired skills not only within the intervention context (e.g., resolving conflicts with peers in the classroom) but also in other settings (e.g., on the playground) or applied to other types of problems (e.g., solving conflicts with adults).
H3.2 Procedures for ensuring application. The conditions under which application outside of the intervention context was achieved; that is, documentation of efforts (or lack thereof) to ensure application to other settings, problems, etc. (e.g., through simulated practice in wide scale application; booster sessions in other contexts).

H3.3 Institutionalization. The extent to which impact on implementers or context is sustained, documented through systematic follow-up. For example, the extent to which teachers continue the application of a classroom-based intervention after the end of the formal intervention program; or the extent to which the school system continues discipline practices established by the intervention program. The conditions under which sustainability has been achieved; that is, documentation of efforts (or lack thereof) to ensure sustained impact (e.g., through booster sessions or follow-up consultation).

H4 Durability/Generalization across persons

Evidence is provided regarding the extent to which outcomes are manifested with participants who are different than the original group of participants for which the intervention was evaluated. An underlying assumption is that each implementation of an intervention in a different context or with a different group constitutes a “different” intervention, given the cultural-contextual nature of interventions.

I. Identifiable Intervention Components

Identifiable components focus on the degree to which the authors of a study have identified unique aspects of the program/intervention that can be linked to statistically significant primary outcomes. Some interventions include multiple components. In these cases, it must be demonstrated that each component is necessary for the primary outcome to occur. For example, if an intervention designed to prevent teen smoking included a knowledge component, a peer component, and a parent component, then the study should demonstrate that all three of these components were necessary for the program to effectively prevent teen smoking. An appropriate method of demonstrating this link would be to design a study that examined the effects of the intervention with all three components, compared to the effects of the same intervention approach with only one or two of the components (e.g., peer and parent components only, knowledge and peer components only, and knowledge and parent components only). Another example of how to examine the effects of specific components might be intervention research where different approaches are compared (e.g., discussion only vs. behavioral training and discussion). As part of this code, reviewers should note the number of program components when that information is available.

Clearly this is a massive undertaking, and the Task Force recognizes that these types of investigations are rare and that the majority of studies will not have addressed this issue. Nevertheless, this process of establishing evidence linking specific intervention components to primary outcomes is critical information that can be used by practitioners and schools in making decisions about whether to implement a particular program, or portions of a particular program. Therefore, this criterion has been included, not because there are a number of studies that necessarily satisfy it currently, but because the Task Force believes this is an important direction for future intervention research.

I1. Evidence for primary outcomes. Rated as follows:

To receive a rating of 3, strong evidence, a study must: (1) demonstrate strong evidence for significant primary outcomes, (2) use a design that allows for an analysis that identifies specific components, and (3) the analysis must provide evidence that all identified intervention components were necessary to produce change in the primary outcomes.

To receive a rating of 2, promising evidence, a study must: (1) demonstrate promising evidence for significant primary outcomes, (2) use a design that allows for an analysis which identifies specific components, and (3) the analysis must provide evidence that at least 50% of the identified intervention components were necessary to produce change in the primary outcomes.

To receive a rating of 1, weak evidence, a study must: (1) demonstrate weak evidence for significant primary
outcomes, (2) use a design that allows for an analysis which identifies specific components, and (3) the analysis must provide evidence that at least 25% of the identified intervention components were necessary to produce change in the primary outcomes.

A rating of 0 indicates that there was no evidence of which components were necessary to produce change.

**ADDITIONAL CRITERIA FOR JUDGING IDENTIFIABLE COMPONENTS** (coded for presence/absence and/or descriptively):

I2. Design allows for analysis of identifiable components.

I3. Total number of components.

I4. Number of components that are linked to the primary outcome.

I5. Documenting essential components. The procedures for insuring delivery of essential components are described in detail, including the procedures for training facilitators, delivery of the intervention, and monitoring the intervention.

I6. Documenting adaptation. The procedures for adapting the intervention to fit the ecology (context, culture, participants) are described in detail, including documentation of the impetus (evidence-based occurrence or feedback) for adaptations and the resulting procedural changes. Each adaptation must be clearly linked to data generated by documentation techniques.

I7. Documenting contextual features. A detailed description of the context within which the intervention occurred is provided. Context refers to the community in which the intervention takes place (e.g., city, neighborhood); the specific site of implementation (e.g., school); and the specific location (e.g., classroom). Contextual variables include geographic location, population (size, demographics), social-cultural variables (e.g., relevant cultural practices or norms, social issues), and resources relevant to the intervention (e.g., necessary equipment and supplies, expenditures).

**J. Implementation Fidelity**

Reviewers will consider program implementation fidelity/integrity. Although there may be slight variations across implementations, data on the program integrity will be critical to determine if the intervention was responsible for the positive outcomes reported by the researchers. Two issues are considered critical here: acceptable adherence and intervention manuals. Reviewers will code whether the authors have confirmed program implementation integrity.

To receive a rating of 3, **strong evidence**, the study must demonstrate strong evidence of acceptable adherence. In addition, evidence should be measured through at least two of the following: ongoing supervision/consultation, coding sessions, or audio/video tapes, and use of a manual. To be considered a “manual” for a rating of 3, information must have been provided to the implementers using either: (1) written materials involving a detailed account of the exact procedures and the sequence in which they are to be used or (2) a formal training session that includes a detailed account of the exact procedures and the sequence in which they are to be used. If the intervention is to be administered in “sessions” or “lessons,” then this information must be provided on a session to session/lesson to lesson basis. Finally, if adaptation occurs to fit varying contexts, there is a description of the procedures for adaptation.

To receive a rating of 2, **promising evidence**, the study must demonstrate evidence of acceptable adherence. In addition, evidence should be measured through at least one of the above criteria and use of a manual. To be considered a “manual” for a rating of 2, information must have been provided to the implementers using either: (1) written materials involving an overview of broad principles and a description of the intervention phases, or (2) a formal or informal training session involving an overview of broad principles and a description of the intervention phases.

To receive a rating of 1, **weak evidence**, the study must demonstrate evidence of acceptable adherence measured
through at least one of the above criteria or use of a manual.

A rating of 0, no evidence, would reflect that nothing was done to ensure implementation fidelity or evidence indicates unacceptable adherence.

J1. Acceptable adherence. The reviewers must be persuaded that the intervention in question was delivered in the study in a manner consistent with the intervention developer’s manual or protocol. Adequate adherence to the intervention developer’s protocol must be shown in all reviewed studies: supporting studies, null studies (i.e., those showing no difference between the candidate intervention and another intervention or control), and refuting studies (i.e., those showing the candidate intervention to be worse than another intervention or control). Acceptable adherence can be met in any one of the following ways:

J1.1 When training, ongoing supervision and case consultation are provided by the intervention development team or another expert in that intervention.

J1.2 When coding of sessions (live, video, or audio) has been done, it must demonstrate that, on average, the majority of authors’ or intervention developers’ criteria for adherence were satisfied.

J1.3 When an intervention is delivered via audio- or videotapes, with or without discussion afterwards.

J 2. Manualization. The candidate intervention should be “manualized” (i.e., accompanied by a clear description of the procedures used) and the studies must be conducted with intervention manuals and/or detailed procedural specification. Manuals may vary in form, depending on the intervention; some manuals may involve a detailed session-by-session account of exact procedures and the sequence in which they are to be used. Others may emphasize broad principles, describe phases of intervention, and offer examples of how the broad principles may be implemented. In some cases (e.g., spelling intervention program), the principles and procedures are straightforward enough that they can be described in the Method section of an article, with no need for a separate manual. In each instance, the test of whether manualization is adequate is whether the intervention procedures are sufficiently well documented that readers know what intervention procedures were tested, and (b) intervention agents / therapists / consultants / instructors and researchers who wish to replicate the intervention could do so faithfully (recognizing that supplemental training may be required for nuances of the intervention to be understood). This documentation must either be present in the article or the documentation (manual and related materials) and must be available to the interested reader / clinician / researcher.

J3. Adaptation. The study describes the principles and procedures for adapting the intervention to fit varying contexts, cultures, and participants while maintaining essential components. Principles of adaptation are evidence-based and derived from repetition of the intervention across multiple contexts, cultures, and samples of participants. The procedural guide advocates repeated evaluation for each application of the intervention.

J4. Implementation Context (Conditions of Implementation). The conditions of implementation are made explicit through documentation involving the use of multiple methods and sources of data (i.e., triangulation) to yield a “rich description” of the conditions under which implementation occurs. Rich description involves in-depth detailing of the procedures and conditions surrounding implementation—including delivery of the intervention, training of facilitators, monitoring implementation, and making necessary adaptation. Implementation context includes characteristics of the implementer(s), adaptations, and the relationship of research to intervention.

J4.1. Characteristics of the implementers. The goal is to provide rich and thorough descriptions of the implementers that goes beyond proxy variables (e.g., age, gender, race, language) and includes deeper and thicker descriptions that include the cultural competence of the interveners and psychological and psychosocial dimensions such as rapport, trust, communication, and ethnic identity.

3 = Provided evidence for high level of similarity to target participants that goes beyond proxy variables (e.g,
rapport, trust, communication, ethnic identity) and cultural competence documented.

2 = Provided evidence for some level of similarity to target participants that goes beyond proxy variables (e.g., rapport, trust, communication, ethnic identity) and cultural competence documented.

1 = Provided no evidence for similarity on these variables, but cultural competence is documented.

0 = Provided no evidence for similarity to target participants and/or cultural competence.

J4.2. Adaptations in implementation. Adaptations made to the intervention are documented in detail.

3 = Detailed account of the implementation and adaptations to fit the context or target population.

2 = Detailed account of the implementation but not of the adaptations to fit the context or target population.

1 = Partial description of the implementation and/or the adaptations to fit the context or target population.

0 = Vague or no account of the implementation.

J4.3. Documents the relationship between the researcher and intervention. The role of the researcher/evaluator in the intervention is described. The participation of the researcher/evaluator in the intervention can vary from passive to active. The level of involvement of the researcher should be described in sufficient detail to permit judgment about the potential impact of the researcher on the intervention. Safeguards for bias of the researcher are specified and can include such practices as triangulation of data sources (e.g., multiple informants or multiple researchers), member checking (confirming findings and interpretations with participants), and researcher’s journal (account of interactions with intervention and participants; personal biases that may influence data collection and interpretation).

3 = Detailed description of the researcher’s level of involvement and safeguards used to minimize the bias of the researcher.

2 = Detailed description of the researcher’s level of involvement, but minimal description of safeguards to minimize the bias of the researcher.

1 = Minimal description of the researcher’s level of involvement and of safeguards to minimize the bias of the researcher.

0 = No information provided.

J4.4. Documents the relationship between the implementers and participants. Description provided regarding procedures and interpersonal processes to establish the relationship for delivering the intervention; for example, nature of the relationship, processes or procedures used to continue pre-existing relationship or establish new relationship, develop and maintain rapport and trust, monitor interpersonal relationships, and facilitate effective working relationships.

3 = Detailed description regarding the interpersonal processes used to establish and maintain the relationship between implementer and participants.

2 = Detailed description of relationship development procedures, but lacking detail on some aspects of the relationship processes.

1 = Provides overview of relationship development procedures and processes, but lack details.

0 = No description of relationship processes provided.

J4.5. Length of Intervention. Reviewers will code the length of the intervention (i.e., how long the intervention was in place) in weeks, months, or years.

J4.6. Intensity of Intervention. Intensity refers to the length of the intervention sessions and the frequency of the sessions. Reviewers will code both length and frequency of intervention sessions.

J4.7. Dosage/Response. Reviewers will note whether there is evidence that length or intensity of intervention was associated with stronger outcomes or better performance on outcome measures.

J4.8. Program Implementer. Reviewers will code who conducted the intervention. Possible choices include:
research staff, school specialty staff (e.g., counselors, psychologists, social workers, special educators, etc.),
teachers, educational assistants, parents, college students, peers, and others. If research staff were only involved
in training intervention implementers, then do not include them as program implementers, unless the intervention
is the training. This support will be documented under Training and Support Resources.

J4.9. Intervention Style or Orientation. Reviewers will note the theoretical underpinnings of the intervention
program, such as behavioral, cognitive-behavioral, experiential, humanistic/interpersonal, psychodynamic/insight
oriented, or combination/other.

J4.10. Cost Analysis Data. Cost analysis data provided (coded yes/no). Whenever authors report data
regarding the cost to implement the program, and/or cost-to-benefit analyses, reviewers will report this
information.

J4.11. Training and Support Resources. What supports were provided if school or other typical staff
implemented the intervention in the study?

J4.11.1. Simple orientation given to change agents.

J4.11.2. Training workshops given (indicate number of workshops, average length of training, and who
conducted the training).

J4.11.3. On-going technical support provided once intervention began, such as phone or face-to-face
consultation, classroom observations, coaching or trouble-shooting after program has begun.

J4.11.4. Program materials must be obtained to conduct intervention.

J4.11.5. Special facilities (extra space) must be obtained.

J4.11.6. Other.

J4.12. Feasibility. Reviewers will rate the level of difficulty in training intervention agents (i.e., high, moderate,
or low) and report and rate the cost involved in training intervention agents (i.e., high, moderate, low) when this
information is available.

K. Replication

Another criterion reviewers will take into account is replication. The number of replications of the program will be
coded. Replication will take into account three issues: the same intervention, the same target problem, and replication
independent of program author group.

K1. Same intervention. Two versions of an intervention program are considered to be the same if (1) the
authors judge the intervention to be essentially the same (declared in the published study, via personal contact, or use
of the same manual/procedures), and (2) the intervention duration is equal, or the shorter version is at least 75% of the
longer version, or a longer intervention is used. If neither, or only one, of these criteria is satisfied, then the two
intervention programs are considered different interventions. Whenever there is a question about whether the
intervention is the same (e.g., manual was revised for group vs. individual administration), the reviewers will attempt to
retrieve all manuals and related materials from the author(s), review them, and come to a consensus on the same
intervention issue.

K2. Same target problem. There must be at least two studies (including independent replications within a
single study) meeting all criteria treating the same target issue/problem, and the same target age/grade group. First,
studies must provide sufficient description of the target issue/problem of the treated group to permit clear identification
of the target issue/problem and target age/grade range to which the intervention is relevant. Second, a study is
considered a replication only when the intervention is applied to a sample of the same target issue/problem and
age/grade range as in the prior study(ies); otherwise, the new study represents a new test with a new target
issue/problem and/or age/grade group. When an intervention is tested with a different target issue/problem or with a
different age/grade group than in previous research, this is considered evidence on a new application of the
intervention, not a replication.

The studies will be grouped by age in the following way: infant (birth to 2 years), preschool (2-5 years),
elementary (6-11 years), secondary (12-17 years), adult (18-65 years), and geriatric (over 65 years). Some examples
of target issue/problem to clarify: if an intervention has two studies supporting its efficacy, one intervening with reading
and the other math, this intervention does not qualify as a replication; if an intervention is supported by two separate
studies, one treating conduct problems and oppositional behavior in children 8-11 years, and the other treating conduct
disorder in children 8-11 years, this intervention may be considered replicated, assuming it meets all other criteria.

K3. Independent evaluation. Reviewers will take into account any relationship between the
evaluator/researcher and the intervention program. Generally, evaluations of an intervention conducted by the program
developers tend to show larger effects than those conducted by independent evaluators. Information pertaining to an
independent evaluation will be coded and noted in all reviews.

To receive a rating of 3, strong evidence, the study must be a replication of the same intervention and target
problem, be implemented by an independent evaluator, and demonstrate similar or better outcomes.

To receive a rating of 2, promising evidence, the study must contain two of the three coding criteria (i.e., same
intervention, same target problem, independent evaluation), and demonstrate similar or better outcomes.

To receive a rating of 1, weak evidence, the study must contain at least one of these coding criteria, and
demonstrate similar or better outcomes.

A rating of 0, no evidence, none of these elements were present, and/or positive outcomes were not
demonstrated.

L. Site of Implementation

Reviewers will note where the intervention took place. In the rating, preference is given to school settings,
although it is recognized that interventions take place in an appropriate field-based site (e.g., home, university clinic,
summer program, outpatient hospital, partial inpatient or day intervention program, inpatient hospital, private practice,
mental health center, or residential intervention facility). One purpose of this information is to indicate whether the
intervention took place in (a) a school setting, (b) a clinical setting where mental health services are routinely delivered
independently of the study (e.g., a community mental health center, HMO, inpatient psychiatric unit), or (c) a setting
specially arranged for research (e.g., university lab clinic, school classroom).

To receive a rating of 3, strong evidence, the study must have been conducted in a public school or an
alternative school, as these are where the majority of children needing intervention are located. If the type of school is
unknown, a rating of 2 is the highest possible rating. In addition, interventions focusing on home-school partnerships
would also receive this rating if the school initiated the intervention in an outreach effort and if the school was a public
or alternative school (e.g., a public school initiated home-based program).

To receive a rating of 2, promising evidence, the study must have been conducted in a private, charter, or
university-affiliated school setting. These settings all contain special variables, that may not exist in public school
settings and that may contribute to intervention effectiveness. In addition, interventions that focus on home-school
partnerships would also receive this rating if the school initiated the intervention in an outreach effort and if the school
was a private, charter, or university-affiliated school setting.

To receive a rating of 1, weak evidence, the intervention was not implemented in a school setting, but could be
implemented with little modification (i.e., a video parent training program that requires videotape vignettes be shown to
a group of parents, followed by a psychologist facilitated discussion could be done within a school context with little modification).

A rating of 0, no evidence, would be an intervention not implemented within a school context and would require major modifications to do so.
Coding Protocol: Group-Based Design

Domain:  □ School- and community-based intervention programs for social and behavioral problems
         □ Academic intervention programs
         □ Family and parent intervention programs
         □ School-wide and classroom-based programs
         □ Comprehensive and coordinated school health services

Name of Coder(s): ______________________________  Date: ____________________________

Full Study Reference in APA format: __________________________________________________________
________________________________________________________________________________________
_______________________________________________________________________________________

Intervention Name (description from study): ____________________________________________________
________________________________________________________________________________________
________________________________________________________________________________________

Study ID Number (Unique Identifier): __________________________________________________________

Type of Publication:  (Check one)

□ Book/Monograph
□ Journal article
□ Book chapter
□ Other (specify):
A. General Study Characteristics

A1. Theoretical basis. Intervention design is based upon a conceptual model that is grounded in theory and applied to the empirical study of the target phenomenon and group.

A1.1. Described theoretical basis and clearly linked to the intervention
A1.2. Described theoretical basis but no connection to the intervention
A1.3. Provided some reference to theory but lacks clear conceptual model and its relation to the intervention
A1.4. Theoretical basis absent

A2. Ecological validity of the empirical-theoretical (conceptual) basis for the target population has been established.

A2.1. Ecological validity of conceptual base established through formative research with target population OR conceptual base is culturally derived
A2.2. Existing theory adapted based on research with target population
A2.3. Existing theory adapted based on empirical evidence with similar or related populations
A2.4. Existing theory used without attention to ecological validity

A3. Researcher perspective of the empirical-theoretical basis for the target population

A3.1. Researcher articulated personal beliefs and relation to the conceptualization of the study
A3.2. Research articulated personal beliefs with vague link to conceptualization
A3.3. Research articulated personal beliefs with no link
A3.4. No discussion of researcher’s beliefs related to study’s conceptualization

A4. Research question guides selection of qualitative, quantitative or mixed methodology of study.

A4.1. Clear links established between research question and methods, and methods are appropriate to the research question.
A4.2. Vague or no links established between research question and methods, but methods are appropriate to the research question
A4.3. Links established between research question and methods, but methods are inappropriate to the research question.
A4.4. No links are established and methods are inappropriate to research question.

A5. Participatory Nature of the Research

A5.1. Participatory (Research Questions): Involvement of participants in research design. Were relevant target group members consulted in the formulation of the research questions?

A5.1.1. Full involvement of target group members (recipients of intervention) in formulating the research questions.
A5.1.2. Involvement of individual representatives of the target group in formulating the research questions.
A5.1.3. Feedback provided by target group members in formulating the research questions.
A5.1.4. No evidence of target group members involvement with formulating research questions.

A5.2. Participatory (Methodology): Involvement of participants in research methods: Were relevant target group members involved in the formulation of the research methods?

A5.2.1. Full involvement of target group members (recipients of intervention) in formulating the research methods.
A5.2.2. Involvement of individual representatives of the target group in formulating the research methods.
A5.2.3. Feedback provided by target group members in formulating the research methods.
A5.2.4. No evidence of target group members involvement with either formulating research methods.

A5.3. Participatory (Implementation): Involvement of participants in implementation of intervention: Were relevant target group members involved in the development and delivery of intervention?

A5.3.1. Full involvement of target group members (recipients of intervention) in formulating and implementing the intervention.
A5.3.2. Partial involvement of target group members in intervention—in either formulating or implementing.
A5.3.3. Involvement of individual representatives of the target group in formulating and/or implementing the intervention.
A5.3.4. No evidence of target group members involvement with formulating or implementing the intervention.

A5.4. Participatory (Negotiated Interpretation): Involvement of participants in interpretation of findings: Were relevant target group members involved in the interpretation of the research findings?

A5.4.1. Full involvement of target group members (recipients of intervention) in interpreting research findings. Interpretation provides authentic representation of participant voices.
A5.4.2. Involvement of individual representatives of the target group in interpreting research findings.
A5.4.3. Feedback provided by target group members about interpretation of results. Representatives are given the opportunity to react to researchers’ interpretations.
A5.4.4. No evidence of target group members involvement with interpreting results.

B. General Design Characteristics

B1. Random assignment designs (if random assignment design, select one of the following)

B1.1. Completely randomized design
B1.2. Randomized block design (between-subjects variation)
B1.3. Randomized block design (within-subjects variation)
B1.4. Randomized hierarchical design

B2. Nonrandomized designs (if nonrandom assignment design, select one of the following)

B2.1. Nonrandomized design
B2.2. Nonrandomized block design (between-participants variation)
B2.3. Nonrandomized block design (within-participants variation)
B2.4. Nonrandomized hierarchical design
B2.5. Optional coding of Quasi-experimental designs (see Appendix C)

B3. Overall confidence of judgment on how participants were assigned (select one of the following)

B3.1. Very low (little basis)
B3.2. Low (guess)
B3.3. Moderate (weak inference)
B3.4. High (strong inference)
B3.5. Very high (explicitly stated)
B3.6. N/A
B3.7. Unknown/unable to code

C. Data Analysis (answer B1 through B6)

C1. Appropriate unit of analysis yes no
C2. Familywise error rate controlled yes no N/A
C3. Sufficiently large $N$  □yes □no
   Statistical Test: __________
   α level: _________________
   ES: _________________
   $N$ required: _______________

C4. Total size of sample (start of the study): ____N
C5. Intervention group sample size: ____N
C6. Control group sample size: ____N

For studies using qualitative data analysis methods, code C7 and C8

C7. Coding
   C7.1 Coding scheme linked to study’s theoretical-empirical basis (select one)  □yes □no
   C7.2 Procedures for ensuring consistency of coding are used (select one)  □yes □no
   Describe procedures: ______________________________________________
   C7.3 Progression from abstract concepts to empirical exemplars is clearly articulated (select one)  □yes □no

C8. Interactive process followed (select one)  □yes □no
   Describe process: ______________________________________________

C9. Rival Interpretations. Evaluates rival cultural, methodological, statistical, or theoretical explanations and hypotheses.
   C9.1 □ Investigated multiple rival hypotheses (e.g., cultural, methodological, statistical, and/or theoretical explanations) and provided empirical evidence
   C9.2 □ Investigated one rival hypothesis (e.g., cultural, methodological, statistical, and/or theoretical explanations) and provided empirical evidence
   C9.3 □ Considered rival explanations without empirical evidence
   C9.4 □ No evidence that rival hypotheses were investigated or considered (examined data only from one perspective)

D. Type of Program (select one)
   D1. □ Universal prevention program
   D2. □ Selective prevention program
   D3. □ Targeted prevention program
   D4. □ Intervention/Treatment
   D5. □ Unknown

E. Stage of the Program (select one)
   E1. □ Model/demonstration programs
   E2. □ Early stage programs
   E3. □ Established/institutionalized programs
   E4. □ Unknown

F. Concurrent or Historical Intervention Exposure (select one)
   F1. □ Current exposure
II. Key Features for Coding Studies and Rating Level of Evidence/ Support

(3=Strong Evidence  2=Promising Evidence  1=Weak Evidence  0=No Evidence)

A. Research Methodology (answer A1 through A5)

A1. Characteristics of the data collector

A1.1 ☐ 3 Provided evidence for high level of similarity to target subjects that goes beyond proxy variables (e.g., rapport, trust, communication, ethnic identity) and cultural competence documented.
A1.2 ☐ 2 Provided evidence for some level of similarity to target subjects that goes beyond proxy variables (e.g., rapport, trust, communication, ethnic identity) and cultural competence documented.
A1.3 ☐ 1 Provided no evidence for similarity on these variables, but cultural competence is documented.
A1.4 ☐ 0 Provided no evidence for similarity to target subjects and/or cultural competence.

A2. Characteristics of Participants

A2.1 ☐ 3 Clear definition of what culture means for groups and measurement of individuals' psychological experience/interpretation of cultural memberships (e.g., acculturation, perceptions of power, oppression)
A2.2 ☐ 2 Clear definition of what culture means for groups without measurement of individuals' psychological experience/interpretation of cultural memberships individuals’ interpretations (Culture is conceptualized but not operationalized)
A2.3 ☐ 1 Equates culture with demographic labels
A2.4 ☐ 0 No attempt to describe, define and/or measure culture

A3. Sample appropriate to research methods. Research methods guide sampling procedures.

A3.1 ☐ 3 Clear links established between research methods and sampling, and sampling is appropriate to the research methods.
A3.2 ☐ 2 Vague or no links established between research methods and sampling, but sampling is appropriate to the research methods.
A3.3 ☐ 1 Links established between research method and sampling, but sampling is inappropriate to the research method.
A3.4 ☐ 0 No links are established and sampling is inappropriate to research methods.

A4. Operationalization. Specifying the link between key abstract constructs (variables) and data collection methods (operations).

A4.1 ☐ 3 Clear links established between constructs and methods, and all key constructs are clearly operationalized.
A4.2 ☐ 2 Some, but not all, key constructs are clearly operationalized.
A4.3 ☐ 1 Vague reference to link between constructs and methods.
A4.4 ☐ 0 No evidence that key constructs are operationalized.

A5. Integration of data from multiple sources, methods, and investigators

A5.1 ☐ 3 Used multiple sources, methods, and investigators.
A5.2 ☐ 2 Used two of the following: multiple sources, multiple methods, multiple investigators
A5.3 ☐ 1 Used one of the following: multiple sources, multiple methods, multiple investigators
A5.4 ☐ 0 No evidence of multiple sources, methods, or investigators
A. OVERALL Rating for Research Methodology (select 0, 1, 2, or 3): □ 3 □ 2 □ 1 □ 0

B. Measurement (answer B1 through B6)

B1. Use of outcome measures that produce reliable scores for the majority of primary outcomes. The table for Primary/Secondary Outcomes Statistically Significant allows for listing separate outcomes and will facilitate decision making regarding measurement (select one of the following)

B1.1 □ Yes  
B1.2 □ No  
B1.3 □ Unknown/unable to code

B2. Multi-method (select one of the following)

B2.1 □ Yes  
B2.2 □ No  
B2.3 □ N/A  
B2.4 □ Unknown/unable to code

B3. Multi-source (select one of the following)

B3.1 □ Yes  
B3.2 □ No  
B3.3 □ N/A  
B3.4 □ Unknown/unable to code

B4. Extent of Engagement—The researchers conduct data collection in a manner that guarantees sufficient scope and depth through prolonged engagement (data collection over a sufficient time period to ensure accuracy of representation) and persistent observation (progressively focused to ensure thorough understanding of consistency and variation), respectively.

B4.1 □ 3 Provided evidence for high level of engagement to ensure deep and accurate representation.  
B4.2 □ 2 Provided evidence for some level of engagement to ensure deep and accurate representation.  
B4.3 □ 1 Provided evidence of minimal level of engagement to ensure deep and accurate representation.  
B4.4 □ 0 Provided no evidence for level of engagement to ensure deep and accurate representation.

B5. Validity of measures reported (select one of the following)

B5.1 □ Yes validated with specific target group  
B5.2 □ In part, validated for general population only  
B5.3 □ No  
B5.4 □ Unknown/unable to code

B6. Cultural Appropriateness of the Measures. In rating this item, consider the following dimensions: meaning, language, dialect, and response format.

B6.1 □ 3 Developed measure for use with target group in study on the basis of empirical evidence (conducted formative research and developed measure)  
B6.2 □ 2 Adapted existing measure for use with target group on the basis of formative research and/or empirical evidence with target group.  
B6.3 □ 1 Developed or adapted measures for use with target group based on empirical evidence with similar or related populations.
B6.4  0  Measure not tailored specifically for target group

B7. Measures of key outcomes are linked to the conceptual model.

  B7.1  3  Clear links established between the conceptual model and key outcome indicators
  B7.2  2  Some, but not all, key outcomes are clearly linked to conceptual model.
  B7.3  1  Vague reference to links between key outcomes and conceptual model
  B7.4  0  No evidence that key outcomes are linked to conceptual model.

B. Overall Rating for Measurement (select 0, 1, 2, or 3):   3   2   1   0

C. Comparison Group

  C1. Type of Comparison Group (select one of the following)

    C1.1  Typical contact
    C1.2  Typical contact (other) specify:
    C1.3  Attention placebo
    C1.4  Intervention elements placebo
    C1.5  Alternative intervention
    C1.6  Pharmacotherapy
    C1.7  No intervention
    C1.8  Wait list/delayed intervention
    C1.9  Minimal contact
    C1.10 Unable to identify comparison group

Rating for Comparison Group (select 0, 1, 2, or 3):   3   2   1   0

  C2. Overall confidence rating in judgment of type of comparison group (select one of the following)

    C2.1  Very low (little basis)
    C2.2  Low (guess)
    C2.3  Moderate (weak inference)
    C2.4  High (strong inference)
    C2.5  Very high (explicitly stated)
    C2.6  Unknown/Unable to code

  C3. Counterbalancing of Change Agents (answer C3.1 to C3.3)

    C3.1  By change agent
    C3.2  Statistical
    C3.3  Other

  C4. Group Equivalence Established (select one of the following)

    C4.1  Random assignment
    C4.2  Post hoc matched set
    C4.3  Statistical matching
    C4.4  Post hoc test for group equivalence

  C5. Equivalent Mortality (answer C5.1 through C5.3)
  C5.1  Low Attrition (less than 20% for Post)
  C5.2  Low Attrition (less than 30% for follow-up)
C5.3  Intent to intervene analysis carried out
Findings __________________

C. OVERALL Rating for Comparison Group (select 0, 1, 2, or 3): □ 3  □ 2  □ 1  □ 0

D. Primary/Secondary Outcomes Are Statistically Significant

D1. Evidence of appropriate statistical analysis for primary outcomes (answer D1.1 through D1.3)

- D1.1  □ Appropriate unit of analysis (rate from previous code)
- D1.2  □ Familywise/experimentwise error rate controlled when applicable (rate from previous code)
- D1.3  □ Sufficiently large N (rate from previous code)

D2. Percentage of primary outcomes that are significant (select one of the following)

- D2.1  □ Significant primary outcomes for at least 75% of the total primary outcome measures for each key construct
- D2.2  □ Significant primary outcomes for between 50% and 74% of the total primary outcome measures for each key construct
- D2.3  □ Significant primary outcomes for between 25% and 49% of the total primary outcome measures for any key construct

Rating for Primary Outcomes Statistically Significant (select 0, 1, 2, or 3): □ 3  □ 2  □ 1  □ 0

D3. Evidence of appropriate statistical analysis for secondary outcomes (answer C3.1 through C3.3)

- D3.1  □ Appropriate unit of analysis
- D3.2  □ Familywise/experimentwise error rate controlled when applicable (rate from previous code)
- D3.3  □ Sufficiently large N (rate from previous code)

D4. Percentage of secondary outcomes that are significant (select one of the following)

- D4.1  □ Significant secondary outcomes for at least 75% of the total secondary outcome measures for each key construct
- D4.2  □ Significant secondary outcomes for between 50% and 74% of the total secondary outcome measures for each key construct
- D4.3  □ Significant secondary outcomes for between 25% and 49% of the total secondary outcome measures for any key construct

Rating for Secondary Outcomes Statistically Significant (select 0, 1, 2, or 3): □ 3  □ 2  □ 1  □ 0

D5. Overall Summary of Questions Investigated

- D5.1 Main effect analyses conducted (select one)  □ yes  □ no
- D5.2 Moderator effect analyses conducted (select one)  □ yes  □ no
  Specify results: _______________________________________________________
- D5.3. Mediator analyses conducted (select one)  □ yes  □ no
  Specify results: _______________________________________________________
D5.4 Unintended outcomes assessed (select one) ☐ yes  ☐ no
Specify results: __________________________________________________________

D6. Cultural Moderator Variables. Investigated effects of cultural moderating variables on the outcome.

D6.1 ☐ 3 Investigated effects of cultural variables, with culture operationalized with reference to acculturation, perceptions of power, oppression.
D6.2 ☐ 2 Investigated effects of cultural variables, with culture operationalized as demographic variables (e.g., race, ethnicity, language)
D6.3 ☐ 1 Interpreted results in terms of culture without formal investigation, but with reference to other research with similar or related cultural groups.
D6.4 ☐ 0 No evidence that moderator effects were investigated and no reference to influence of culture in interpretation of findings.

D. OVERALL Rating for Primary/Secondary Outcomes (select 0, 1, 2, or 3): ☐ 3  ☐ 2  ☐ 1  ☐ 0
D. Primary/Secondary Outcomes Statistically Significant (only list $p \leq .05$)

(List primary outcomes first in alphabetical order, followed by secondary outcomes in alphabetical order)

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Primary vs. Secondary</th>
<th>Cultural Relevance</th>
<th>Who Changed</th>
<th>What Changed</th>
<th>Source</th>
<th>Treatment Information</th>
<th>Outcome Measure Used</th>
<th>Reliability</th>
<th>ES $(1-\beta)$</th>
</tr>
</thead>
</table>
Null Findings/Negative Outcomes Associated with the Intervention (listed alphabetically by outcome)

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Primary vs. Secondary</th>
<th>Cultural Relevance</th>
<th>Who Was Targeted for Change</th>
<th>What Was Targeted for Change</th>
<th>Source</th>
<th>Note null/negative outcomes</th>
<th>Outcome Measure Used</th>
<th>Reliability</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Teacher</td>
<td>Attitude</td>
<td>Parent Report</td>
<td>Teacher Report</td>
<td>Observation</td>
<td>Test</td>
<td>Other</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Parent/sign. Adult</td>
<td>Knowledge</td>
<td>Test</td>
<td>Other</td>
<td>Unknown</td>
<td>Test</td>
<td>Other</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Ecology</td>
<td>Other</td>
<td>Other</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Other</td>
<td>Unknown</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome #2</td>
<td>Unknown</td>
<td></td>
<td>Child</td>
<td>Behavior</td>
<td>Self Report</td>
<td>Parent Report</td>
<td>Teacher Report</td>
<td>Observation</td>
<td>Test</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Teacher</td>
<td>Attitude</td>
<td>Parent Report</td>
<td>Teacher Report</td>
<td>Observation</td>
<td>Test</td>
<td>Other</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Parent/sign. Adult</td>
<td>Knowledge</td>
<td>Test</td>
<td>Other</td>
<td>Unknown</td>
<td>Test</td>
<td>Other</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Ecology</td>
<td>Other</td>
<td>Other</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Other</td>
<td>Unknown</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Teacher</td>
<td>Attitude</td>
<td>Parent Report</td>
<td>Teacher Report</td>
<td>Observation</td>
<td>Test</td>
<td>Other</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Parent/sign. Adult</td>
<td>Knowledge</td>
<td>Test</td>
<td>Other</td>
<td>Unknown</td>
<td>Test</td>
<td>Other</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Ecology</td>
<td>Other</td>
<td>Other</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Other</td>
<td>Unknown</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Teacher</td>
<td>Attitude</td>
<td>Parent Report</td>
<td>Teacher Report</td>
<td>Observation</td>
<td>Test</td>
<td>Other</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Parent/sign. Adult</td>
<td>Knowledge</td>
<td>Test</td>
<td>Other</td>
<td>Unknown</td>
<td>Test</td>
<td>Other</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Ecology</td>
<td>Other</td>
<td>Other</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Other</td>
<td>Unknown</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Teacher</td>
<td>Attitude</td>
<td>Parent Report</td>
<td>Teacher Report</td>
<td>Observation</td>
<td>Test</td>
<td>Other</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Parent/sign. Adult</td>
<td>Knowledge</td>
<td>Test</td>
<td>Other</td>
<td>Unknown</td>
<td>Test</td>
<td>Other</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Ecology</td>
<td>Other</td>
<td>Other</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td>Other</td>
<td>Unknown</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type of Data Effect Size is Based On</td>
<td>Confidence Rating in ES Computation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>-------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(check all that apply)</td>
<td>(select one of the following)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>☐ Means and SDs</td>
<td>☐ Highly estimated (e.g., only have ( N ) ( p ) value)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>☐ ( t )-value or ( F )-value</td>
<td>☐ Moderate estimation (e.g., have complex but complete statistics)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>☐ Chi-square (( df = 1 ))</td>
<td>☐ Some estimation (e.g., unconventional statistics that require conversion)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>☐ Frequencies or proportions (dichotomous)</td>
<td>☐ Slight estimation (e.g., use significance testing statistics rather than descriptives)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>☐ Frequencies or proportions (polytomous)</td>
<td>☐ No estimation (e.g., all descriptive data is present)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>☐ Other (specify):</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>☐ Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
E. Cultural Significance

E.1 3 Conducted formative research or provided empirical evidence with the target population to show relevance of key outcome variables and indicators to target population
E.2 2 Cited empirical evidence for relevance of key outcome variables and indicators to target population
E.3 1 Cited empirical evidence with similar or related populations for relevance of key outcome variables and indicators to target population
E.4 0 No empirical evidence cited for relevance of outcome variables and indicators to population.

E. OVERALL Rating for Cultural Significance (select 0, 1, 2, or 3): 3 2 1 0

F. Educational/Clinical Significance

<table>
<thead>
<tr>
<th>Outcome Variables:</th>
<th>Pretest</th>
<th>Posttest</th>
<th>Follow Up</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1. Categorical Diagnosis Data</td>
<td>Diagnostic information regarding inclusion into the study presented: □ Yes □ No □ Unknown</td>
<td>Positive change in diagnostic criteria from pre to posttest: □ Yes □ No □ Unknown</td>
<td>Positive change in diagnostic criteria from posttest to follow up: □ Yes □ No □ Unknown</td>
</tr>
<tr>
<td>F2. Outcome Assessed via continuous Variables</td>
<td></td>
<td>Positive change in percentage of participants showing clinical improvement from pre to posttest: □ Yes □ No □ Unknown</td>
<td>Positive change in percentage of participants showing clinical improvement from posttest to follow up: □ Yes □ No □ Unknown</td>
</tr>
<tr>
<td>F3. Subjective Evaluation: The importance of behavior change is evaluated by individuals in direct contact with the participant.</td>
<td>Importance of behavior change is evaluated: □ Yes □ No □ Unknown</td>
<td>Importance of behavior change from pre to posttest is evaluated positively by individuals in direct contact with the participant: □ Yes □ No □ Unknown</td>
<td>Importance of behavior change from posttest to follow up is evaluated positively by individuals in direct contact with the participant: □ Yes □ No □ Unknown</td>
</tr>
<tr>
<td>F4. Social Comparison: Behavior of participant at pre, post, and follow up is compared to normative data (e.g., a typical peer).</td>
<td>Participant’s behavior is compared to normative data □ Yes □ No □ Unknown</td>
<td>Participant’s behavior has improved from pre to posttest when compared to normative data: □ Yes □ No □ Unknown</td>
<td>Participant’s behavior has improved from posttest to follow up when compared to normative data: □ Yes □ No □ Unknown</td>
</tr>
</tbody>
</table>

F. Overall Rating for Educational/Clinical Significance (select 0, 1, 2, or 3): 3 2 1 0
G. External Validity Indicators

G1. Sampling Procedures (Answer G1.1 through G1.4)

G1.1 Sampling procedures described in detail
   G1.1.1. 1 Yes
   G1.1.2. 0 No (incomplete or no evidence)

   G1.2.1. 1 Yes
   Specify: ________________________________
   G1.2.2. 0 No (incomplete or no evidence)

G1.3. Rationale for sample size specified
   G1.3.1. 1 Yes
   Specify: ________________________________
   G1.3.2. 0 No (incomplete or no evidence)

G1.4. Evidence provided that sample represents target population
   G1.4.1. 1 Yes
   G1.4.2. 0 No (incomplete or no evidence)

G1.5 Recruitment procedures congruent with target cultural group. Researcher used culturally appropriate ways/methods to contact, recruit, inform, and maintain participation.
   G1.5.1. 1 Yes
   G1.5.2. 0 No (inadequate description or no evidence)

G1.6. Inclusion/exclusion criteria specified □yes □no

G1.7. Inclusion/exclusion criteria similar to school practice □yes □no

G1.8. Specified criteria related to concern □yes □no

G1. Rating for Sampling (select 0, 1, 2, or 3): □ 3 □ 2 □ 1 □ 0
## G2. Participant Characteristics Specified for Treatment and Control Group

<table>
<thead>
<tr>
<th>Participants from Treatment Group</th>
<th>Grade/age</th>
<th>Gender</th>
<th>Ethnicity or Multiethnic Id</th>
<th>Ethnic Identity</th>
<th>Race(s)</th>
<th>Acculturation</th>
<th>Primary Language</th>
<th>SES</th>
<th>Family Structure</th>
<th>Locale</th>
<th>Disability</th>
<th>Functional Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child/Student</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent/caregiver</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child/Student</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent/caregiver</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child/Student</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent/caregiver</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child/Student</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent/caregiver</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child/Student</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent/caregiver</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child/Student</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent/caregiver</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participants from Control Group</td>
<td>Grade/age</td>
<td>Gender</td>
<td>Ethnicity or Multiethnic</td>
<td>Ethnic Identity</td>
<td>Race(s)</td>
<td>Acculturation</td>
<td>Primary Language</td>
<td>SES</td>
<td>Family Structure</td>
<td>Locale</td>
<td>Disability</td>
<td>Functional Descriptors</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-----------</td>
<td>--------</td>
<td>--------------------------</td>
<td>----------------</td>
<td>---------</td>
<td>--------------</td>
<td>-----------------</td>
<td>-----</td>
<td>-----------------</td>
<td>--------</td>
<td>------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>Child/Student</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent/caregiver</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child/Student</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent/caregiver</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child/Student</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent/caregiver</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child/Student</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent/caregiver</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child/Student</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent/caregiver</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child/Student</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent/caregiver</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child/Student</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent/caregiver</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child/Student</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent/caregiver</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
G3. Adequately reported characteristics of participants/sample. Adequate level of detail in description of participants.

G3.1. □ 1 Yes
G3.2 □ 0 No (Incomplete or no evidence)

G4. Details are provided regarding variables that:

G4.1 Have differential relevance for intended outcomes □ yes □ no
Specify: ____________________________

G4.2 Have relevance to inclusion criteria □ yes □ no
Specify: ____________________________

G5. Transferability of the intervention.

G5.1. □ 3 Complete and detailed description of the context within which the intervention occurs.
G5.2 □ 2 Detailed description of some but not all contextual components
G5.3 □ 1 Provides overview of contextual components but lack details
G5.4 □ 0 No description of context

G6. Participant perceptions of benefits of intervention (treatment group)

G6.1. □ 3 Provided evidence of perceived benefits from the intervention for all participant groups.
G6.2 □ 2 Provided evidence of perceived benefits from the intervention for some participant groups.
G6.3 □ 1 Provided evidence that participants did not perceive benefits from the intervention.
G6.4 □ 0 Did not investigate participants’ perceptions of benefits.

G. OVERALL Rating for External Validity (select 0, 1, 2, or 3): □ 3 □ 2 □ 1 □ 0
<table>
<thead>
<tr>
<th>Participants from Treatment Group</th>
<th>Results (What person reported to have gained from participation in program)</th>
<th>General Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child/Student</td>
<td></td>
<td>Participants reported benefiting overall from the intervention</td>
</tr>
<tr>
<td>Parent/caregiver</td>
<td></td>
<td>Participants reported not benefiting overall from the intervention</td>
</tr>
<tr>
<td>Teacher</td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child/Student</td>
<td></td>
<td>Participants reported benefiting overall from the intervention</td>
</tr>
<tr>
<td>Parent/caregiver</td>
<td></td>
<td>Participants reported not benefiting overall from the intervention</td>
</tr>
<tr>
<td>Teacher</td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child/Student</td>
<td></td>
<td>Participants reported benefiting overall from the intervention</td>
</tr>
<tr>
<td>Parent/caregiver</td>
<td></td>
<td>Participants reported not benefiting overall from the intervention</td>
</tr>
<tr>
<td>Teacher</td>
<td></td>
<td></td>
</tr>
<tr>
<td>School</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
H. Durability/Generalization of Intervention and Outcomes

H1. Follow-up assessment

H1.1 Timing of follow up assessment: □ yes □ no
Specify____________________

H1.2 Number of participants included in the follow up assessment: □ yes □ no
Specify____________________

H1.3 Consistency of assessment method used: □ yes □ no
Specify____________________

H1.4 Follow-up addresses institutionalization and/or sustainability of intervention efforts:
□ yes □ no Specify: ________________________________

Overall Rating for Follow-up Assessment (select 0, 1, 2, or 3): □ 3 □ 2 □ 1 □ 0

H2. Durability/Generalization over time

H2.1.1 Evidence is provided regarding the sustainability of outcomes after intervention is terminated □ yes □ no
Specify:__________________________

H2.1.2 Procedures for maintaining outcomes are specified □ yes □ no
Specify:__________________________

H3. Durability/Generalization across settings

H3.1 Evidence is provided regarding the extent to which outcomes are manifested in contexts that are different from the intervention context □ yes □ no
Specify:__________________________

H3.2 Documentation of efforts to ensure application of intervention to other settings □ yes □ no
Specify:__________________________

H3.3 Impact on implementers or context is sustained □ yes □ no
Specify:__________________________

H4. Generalization across persons
Evidence is provided regarding the degree to which outcomes are manifested with participants who are different than the original group of participants for with the intervention was evaluated: yes no

Specify: ________________________________

G. OVERALL Rating Durability/Generalization (select 0, 1, 2, or 3): 3 2 1 0

I. Identifiable Intervention Components (answer I1 through I7)

I1. Overall Rating for Identifiable Components: 3 2 1 0

I2. Design allows for analysis of identifiable components (select one) yes no

I3. Total number of components: ______ N

I4. Number of components linked to primary outcomes: ______ N

I5. Clear documentation of essential components (select one) yes no

Specify: ________________________________

I6. Procedures for adapting the intervention are described in detail (select one) yes no

Specify: ________________________________

I7. Contextual features of the intervention are documented (select one) yes no

Specify: ________________________________

G. OVERALL Rating of Identifiable Intervention Components (select 0, 1, 2, or 3): 3 2 1 0

J. Implementation Fidelity

J1. Evidence of Acceptable Adherence (answer J1.1 through J1.3)

J1.1 □ Ongoing supervision/consultation
J1.2 □ Coding intervention sessions/lessons or procedures
J1.3 □ Audio/video tape implementation (select J1.3.1 or J1.3.2):
    J1.3.1 □ Entire intervention
    J1.3.2 □ Part of intervention

J2. Manualization (select all that apply)

J2.1 □ Written material involving a detailed account of the exact procedures and the sequence in which they are to be used

J2.2 □ Formal training session that includes a detailed account of the exact procedures and the sequence in which they are to be used
J2.3  □ Written material involving an overview of broad principles and a description of the intervention phases

J2.4  □ Formal or informal training session involving an overview of broad principles and a description of the intervention phases

J3. Adaptation procedures are specified (select one)  □ yes □ no □ unknown

**Rating for Implementation Fidelity** (select 0, 1, 2, or 3):  □ 3 □ 2 □ 1 □ 0

J4. Implementation Context (Conditions of Implementation)

J4.1. Characteristics of the Implementer

J4.1.1  □ 3 Provided evidence for high level of similarity to target subjects that goes beyond proxy variables (e.g., rapport, trust, communication, ethnic identity) **and** cultural competence documented.

J4.1.2  □ 2 Provided evidence for some level of similarity to target subjects that goes beyond proxy variables (e.g., rapport, trust, communication, ethnic identity) and cultural competence documented.

J4.1.3  □ 1 Provided no evidence for similarity on these variables, but cultural competence is documented.

J4.1.4  □ 0 Provided no evidence for similarity to target subjects and/or cultural competence.

J4.2 Adaptations in Implementation

J4.2.1  □ 3 Detailed account of the implementation and adaptations to fit the context or target population

J4.2.2  □ 2 Detailed account of the implementation but not of the adaptations to fit the context or target population

J4.2.3  □ 1 Partial description of the implementation and/or the adaptations to fit the context or target population

J4.2.4  □ 0 Vague or no account of the implementation

J4.3 Relationship of Researcher to Intervention

J4.3.1.  □ 3 Detailed description of the researcher’s level of involvement and safeguards used to minimize the bias of the researcher.

J4.3.2.  □ 2 Detailed description of the researcher’s level of involvement, but minimal description of safeguards to minimize the bias of the researcher.

J4.3.3.  □ 1 Minimal description of the researcher’s level of involvement and of safeguards to minimize the bias of the researcher.

J4.3.4  □ 0 No information provided

J4.4 Relationship of Implementer/to Participants

J4.4.1.  □ 3 Detailed description regarding the interpersonal processes used to establish and maintain the relationship between implementer and participants.

J4.4.2.  □ 2 Detailed description of relationship development procedures, but lacking detail on some aspects of the relationship processes.

J4.4.3.  □ 1 Provides overview of relationship development procedures and processes, but lack details

J4.4.4.  □ 0 No description of relationship processes provided.

J4.5 Length of Intervention (select J4.5.1 or J4.5.2)
J4.5.1. □ Unknown/insufficient information provided

J4.5.2. □ Information provided (if information is provided, specify one of the following:)

J4.5. 2.1. weeks _____ N

J4.5. 2..2. months _____ N

J4.5. 2.3. years _____ N

J4.5. 2.4 other _____ N

J4.6   Intensity/dosage of Intervention (select J4.6.1 or J4.6.2)

J4.6.1. □ Unknown/insufficient information provided

J4.6.2. □ Information provided (if information is provided, specify both of the following:)

J4.6.2.1 length of intervention session _____ N

J4.6.2.2 frequency of intervention session _____ N

J4.7 Dosage Response (select J4.7.1 or J.7.2)

J4.7.1. □ Unknown/insufficient information provided

J4.7.2. □ Information provided (if information is provided, answer J4.7.2.1)

J4.7.2.1 Describe positive outcomes associated with higher dosage: ______________________

J4.8 Program Implementer (select all that apply)

J4.8.1. □ Research Staff

J4.8.2. □ School Specialty Staff

J4.8.3. □ Teachers

J4.8.4. □ Educational Assistants

J4.8.5. □ Parents

J4.8.6. □ College Students

J4.8.7. □ Peers

J4.8.8. □ Other

J4.8.9. □ Unknown/insufficient information provided

J4.9 Intervention Style or Orientation (select all that apply)

J4.9.1. □ Behavioral

J4.9.2. □ Cognitive-behavioral

J4.9.3. □ Experiential

J4.9.4. □ Humanistic/interpersonal

J4.9.5. □ Psychodynamic/insight oriented

J4.9.6. □ other (specify): ______________________

J4.9.7. □ Unknown/insufficient information provided

J4.10   Cost Analysis Data (select G1 or G2)

J4.10.1. □ Unknown/insufficient information provided

J4.10.2. □ Information provided (if information is provided, answer H2.1)
J4.10.2.1 Estimated Cost of Implementation: __________________________

J4.11 Training and Support Resources (select all that apply)

J4.11.1. □ Simple orientation given to change agents
J4.11.2. □ Training workshops conducted

# of Workshops provided ______
Average length of training ______
Who conducted training (select all that apply)
J4.11.2.1 □ Project Director
J4.11.2.2 □ Graduate/project assistants
J4.11.2.3 □ Other (please specify): ______
J4.11.2.3 □ Unknown

J4.11.3. □ Ongoing technical support
J4.11.4. □ Program materials obtained
J4.11.5. □ Special Facilities
J4.11.6. □ Other (specify):

J4.12 Feasibility

J4.12.1 Level of difficulty in training intervention agents (select one of the following)
J4.12.1.1 □ High
J4.12.1.2 □ Moderate
J4.12.1.3 □ Low
J4.12.1.4 □ Unknown

J4.12.2. Cost to train intervention agents (specify if known): ______________________

J4.12.3. Rating of cost to train intervention agents (select one of the following)
J4.12.3.1 □ High
J4.12.3.2 □ Moderate
J4.12.3.3 □ Low
J4.12.3.4 □ Unknown

J. OVERALL Rating for Identifiable Intervention Components (select 0, 1, 2, or 3): 3 2 1 0

K. Replication (answer K1, K2, K3, and K4)

K1. □ Same Intervention
K2. □ Same Target Problem
K3. □ Independent evaluation

K. OVERALL Rating for Replication (select 0, 1, 2, or 3): 3 2 1 0

L. Site of Implementation

L1. School (if school is the site, select one of the following options)
L1.1 [ ] Public
L1.2 [ ] Private
L1.3 [ ] Charter
L1.4 [ ] University Affiliated
L1.5 [ ] Alternative
L1.6 [ ] Not specified/unknown

L2. Non School Site (if it is a non school site, select one of the following options)

L2.1 [ ] Home
L2.2 [ ] University Clinic
L2.3 [ ] Summer Program
L2.4 [ ] Outpatient Hospital
L2.5 [ ] Partial inpatient/day Intervention Program
L2.6 [ ] Inpatient Hospital
L2.7 [ ] Private Practice
L2.8 [ ] Mental Health Center
L2.9 [ ] Residential Treatment Facility
L2.10 [ ] Other (specify): ____________________________
L2.11 [ ] Unknown/insufficient information provided

L. OVERALL Rating for Site of Implementation (select 0, 1, 2, or 3):  [ ] 3  [ ] 2  [ ] 1  [ ] 0
## Summary of Evidence for Group-Based Design Studies

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Overall Evidence Rating</th>
<th>Description of Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NNR = No numerical rating or 0 - 3</td>
<td>Strong Promising Weak No/limited evidence or Descriptive ratings</td>
</tr>
</tbody>
</table>

### General Characteristics

- General Study Characteristics
- General Design Characteristics
- Data Analysis
- Type of Program
- Stage of Program
- Concurrent/Historical Intervention Exposure

### Key Features

- Research Methodology
- Measurement
- Comparison Group
- Primary/Secondary Outcomes are Statistically Significant
- Cultural Significance
- Educational/Clinical Significance
- External Validity Indicators
- Durability/Generalization
- Identifiable Intervention Components
- Implementation Fidelity
- Replication
- Site of Implementation