WWC EVIDENCE REVIEW PROTOCOL:
Interventions for Children Classified as Having an Emotional Disturbance
(Version 2.0)

Topic Area Focus

This What Works Clearinghouse (WWC) review focuses on interventions designed to meet the academic, behavioral, social, and emotional needs of K–12 students with an emotional disturbance, as well as students formally described as at risk of being classified as having an emotional disturbance. The emotional disturbance classification encompasses several psychiatric, behavioral, and emotional concerns; related interventions range from behavior modification, psychiatric medication, and psychotherapy to nutritional changes.

The intended audience for this review, however, is educators working in K–12 settings who have an interest in serving students with an emotional disturbance or students at risk for classification. Hence, only those interventions with a school-based component will be reviewed. To clarify, the review will focus on any intervention that is delivered, at least in part, in a K–12 school site as long as school-based staff members have a substantial role in its delivery using resources at their disposal. This review will also include interventions with community and home-based components, as long as school staff members have responsibility for directly delivering a treatment component. This treatment often will involve a behavior plan, school-based counseling, and/or consultation. By contrast, simply collecting data for others to interpret outside of a school-based setting would suggest that the intervention does not have a substantial school-based component. This focus precludes investigation of, for example, medication trials and therapy studies outside of a school setting.

Systematic reviews of evidence in this topic area address the following questions:

- Which interventions are effective in addressing the behavioral, academic, emotional, and/or social needs of students classified with an emotional disturbance or at risk for classification?

- Are some interventions more effective than others for certain types of students, particularly those who are formally defined as at risk for being classified?

Intervention-level reports will describe subgroup analyses (based on characteristics such as severity of disability, age, gender, school level, and setting in which the intervention was delivered); immediate versus sustained effects; and whether interventions entailed single or multiple elements. We will also describe authors' attempts to establish intervention fidelity should they discuss the matter, using their descriptions and conclusions while indicating clearly that we are making no judgment about these data. Additional information is provided throughout this protocol. Individual reports will be released on a periodic basis.

Key Definitions

Emotional Disturbance

Emotional disturbance is defined in the Federal Register as a "condition exhibiting one or more of the following characteristics over a long period of time and to a marked degree that adversely affects a child's educational performance:

(A) An inability to learn that cannot be explained by intellectual, sensory, or health factors.
(B) An inability to build or maintain satisfactory interpersonal relationships with peers and teachers.
(C) Inappropriate types of behavior or feelings under normal circumstances.
(D) A general pervasive mood of unhappiness or depression.
(E) A tendency to develop physical symptoms or fears associated with personal or school problems.

Emotional disturbance includes schizophrenia, but does not apply to children who are socially maladjusted, unless it is determined that they have an emotional disturbance" (Assistance to States for the Education of Children with Disabilities and Preschool Grants for Children with Disabilities, 2006).

It is noteworthy that *emotional or behavioral disorder* (EBD) is a competing term in the literature, and the search for this review may identify studies that refer to samples as having a *serious emotional disturbance, emotional handicap, behavioral disorder*, or *serious behavior disorder*. States also vary in how they define *emotional disturbance* (hereafter abbreviated as ED), using differing inclusion criteria, and some may apply nuanced language, such as "children with ED" in lieu of "ED children." For the purposes of this review, these are interchangeable terms even though there are substantive reasons for competing definitions.[1]

At Risk for ED Classification and Prevention Programs

Students may exhibit severe and persistent problems that fall within the ED spectrum but are not yet formally classified. These students are deemed to be at risk for an ED classification, and this status can be used to qualify a sample for review. At-risk status should be supported by behavioral rating scales, teacher reports, school records, or presence of a diagnosed psychiatric disorder (more on this below). The key inclusion criterion is that study authors must make the case that the sample is at risk for being classified with an ED because of the presence of an emotional or behavioral concern. An implication of including at-risk samples is that prevention programs may be reviewed. Prevention programs are increasingly becoming part of special education practice, and related policy has recognized and funded such programs.

Including at-risk samples yields a complex issue because this special education classification has ambiguous boundaries. Consider the federal legislation's social maladjustment exemption. There appear to be no systematic criteria for separating out children with such maladjustment from

[1] The primary reason for the varying definitions is that the IDEA version is a matter of public policy and not necessarily research. The definition draws heavily from Bower's (1981) definition of emotionally handicapped children and so arguably does not reflect current practice (Forness & Knitzer, 1992; Kauffman & Landrum, 2009). For example, the IDEA version makes no mention of preventive approaches, even though they represent a core feature of practices and research funded by public dollars (e.g., the current focus on Response-to-Intervention and School Wide Behavioral Supports). Because the purpose of this endeavor is to summarize intervention impacts for children with an ED and not to enter the controversy around defining ED, all of the previously mentioned terms are accepted as a basic means for qualifying a study for review.

students with an ED or for that matter even defining the term (Assistance to States for the Education of Children with Disabilities and Preschool Grants for Children with Disabilities, 2006). Furthermore, children who will probably never require special education services may still at times present difficulties used for classification purposes (such as the inability to build or maintain satisfactory interpersonal relationships with peers and teachers or inappropriate types of behavior or feelings under normal circumstances). As Kauffman and Landrum (2009) note, such children can present the same behaviors as students with an ED, with the main differences being persistence and pervasiveness of behavior. This makes the inclusion of studies dealing with students who are at risk potentially problematic because one might include practically any K–12 intervention with an emotional, behavioral, or social component. This concern is handled below in the section describing general inclusion criteria.

Psychiatric Disorders

The ED classification encompasses many psychiatric disorders, although the presence of such conditions is not necessary for special education classification. For the purposes of this review, a *psychiatric disorder* is a formally diagnosed mental health condition that is used as a basis for an ED classification or for establishing that a student is at risk for being classified. Examples include, but are not limited to, obsessive compulsive disorder, depression, selected mutism, schizophrenia, and somatic disorders. Additional examples of relevant diagnoses are presented below in the keyword search section. The existence of one of these psychiatric disorders is a qualifying sample characteristic when it comes to screening studies for inclusion in the review pool.

An important exception is attention deficit hyperactivity disorder (ADHD). Children exhibiting sufficient severity and persistence of inattention, hyperactivity, and impulsivity to result in impairment in several settings are typically diagnosed with ADHD. The literature on ADHD is both expansive and in our judgment distinct from the related literature. Furthermore, children with ADHD often are served under the Americans with Disabilities Act (i.e., 504 plans) and not special education. If sufficiently served via a 504 plan, they may be at little risk for being classified with an ED. In sum, studies of ADHD interventions are ineligible unless the sample of children is also classified as having an emotional disturbance or described as at risk of being classified with an emotional disturbance. Interventions that focus on the needs of children with ADHD will be investigated in a separate review.

Autism is another disorder that will not be addressed in this topic area but will be investigated in a separate review. Autism is a developmental disorder characterized by pervasive impairment of communication abilities. Like ADHD, autism research is distinct from the ED field. More important, autism has its own special education category. As is the case with children with ADHD, any sample diagnosed as autistic *and* having an ED classification will be included, irrespective of the reason behind the classification. It is possible that an autism spectrum disorder known as Asperger's syndrome may lead to an ED classification, so this diagnosis can be used to help meet sample criteria for this review. Finally, the presence of eating disorders (alone) is not sufficient for the purposes of this review as related concerns will rarely require services provided to children with an ED classification. Again, the disorders in the keyword list represent qualifying criteria.

Self-Contained Settings

Most children with a special education classification are educated in general education settings because of the concept of *least restrictive environment* (briefly, the idea that students with disabilities should be educated in general education or inclusive classrooms to the maximum extent possible). Self-contained settings entail pull-out instruction for part or all of the school day and may go so far as to include specialized schools, residential arrangements, and hospitals. As long as a substantial part of the intervention is delivered via K–12 school staff (in a school building), it is eligible for review. It is probable that many single-case designs may take place in self-contained settings.

GENERAL INCLUSION CRITERIA

Populations to be Included

The population of interest includes K–12 special education students between ages 5 and 21 who are classified with an ED or are explicitly described as at risk for being classified. In studies that provide aggregated data for both preschool and kindergarten children who received the intervention, and disaggregated data are not available, the review will include the study if at least 50% of the children are in kindergarten. Because we will include at-risk samples, we require that authors formally indicate that the intervention is intended for classification prevention. At-risk samples should be described as exhibiting at-risk behavioral/emotional criteria pertaining to EBD spectrum concerns (based on teacher reports, scores on baseline measures, etc.). The term *at risk* need not be literally used as long as there is evidence that the sample of students has behavioral and/or emotional problems and the intent is to address these problems before they get worse. Our definition of at risk *does not* include students who are described as at risk simply because of geographic location, race/gender, socioeconomic status, or academic standing.

Students may be attending elementary, middle, or high school at any type of school (public, private, parochial, etc.). Samples with multiple classifications or diagnoses will be included in the review as long as ED or at-risk status (as identified by study authors) is present. This review will exclude students with needs so severe that they are not served by typical schools (e.g., students in hospitals, residential treatment programs, or schools that serve only students with an ED or incarcerated youth).

If the sample includes children with an ED (including those described as at risk for classification) *and* those without an ED, reviewers will determine if ED subsample comparisons were conducted.[2] If the comparisons are mentioned in the report and are available (either via the report or author query), the ED subsample will be included in the review. If such comparisons are not conducted in the context of a mixed sample, a study's aggregated sample can still be reviewed if (1) the intervention has a clear focus on meeting the needs of students with an ED (including prevention) and (2) at least 50% of the sample is described as at risk for or classified as having an ED. Recall that the purpose of this review is to promote a focus on studies that investigate impacts of interventions for children who exhibit severe enough behavior problems that they are at risk for special education classification and to avoid including studies for a broader population of students, such as those that address schoolwide disciplinary practices.

Children must reside in the United States, its territories, or tribal entities. Both children who speak English and those who are or are English Language Learners will be included in reviews.

When results are available for subgroups of children defined by the following characteristics, they will be included in an appendix to the intervention report: age, gender, socioeconomic status, race/ethnicity, English Language Learner status, formally classified versus at risk, and severity of disability.

When results are available for subgroups of settings based on the following characteristics, they will be documented in an appendix to the intervention report: location (urban, suburban, rural); setting (school or home-based); school level (elementary, middle, high school); intervention setting (segregated setting such as self-contained classrooms or inclusive setting); staff

[2] Assume moving forward that reference to ED also includes students who are at risk for ED, unless otherwise specified.

credentials, education, qualifications, or training (such as certification or years of experience); and whether an intervention was delivered in the context of broader wraparound services.

When the target population of an intervention overlaps with those of the Character Education and/or Drop Out Prevention WWC reviews, the principal investigator (PI) will consult with the PI of these topic areas to determine which team(s) should review the study.

Types of Interventions to be Included

The overall goal of the review is to inform educators about impacts of interventions that they can deliver to students with an ED. Thus, this review encompasses interventions that have a school-based component, delivered either as part of a school's set of services or in the students' homes under the direction of school staff. The key criterion is that interventions must *target* the needs of students who are *explicitly classified or at risk of being classified as having an ED*. The term *needs* is purposefully broad. The intervention can focus on behavioral, emotional, social, and/or academic outcomes.

Interventions that are delivered as part of wraparound plans, which entail coordinated services across community systems, are eligible for review as long as there is a clear school-based component. Programs with home-based components, such as parent delivery of a behavioral intervention plan, must be managed by school staff, meaning that outcome data are collected by school staff, and there is evidence that school staff have direct responsibility for implementation. Again, interventions delivered primarily in nonschool settings such as residential treatment programs, hospitals, or prisons are ineligible for review.

To be reviewed, interventions must be replicable. That is, the intervention must be branded, or the following elements of the intervention must be documented: target population, characteristics of settings in which it was implemented, specification of key features or components of the intervention, characteristics of the intervention duration and intensity, and staff training required to implement the intervention.

Reviews will document reported intervention fidelity. That is, we will describe authors' attempts to establish fidelity but will ignore the issue if it is not discussed in original study reports. Because we are summarizing only study authors' statements of how fidelity was assessed and the degree to which fidelity was achieved, we will not make judgments about the quality of related data or related analyses and interpretations. Furthermore, we will indicate whether the study was identified (by study authors) as efficacy or effectiveness research. Again, we will not judge authors' statements but will simply record how they classify the research should they discuss the issue. On a related point, we will discuss if interventions were carried out by typical school staff or researchers.

The types of interventions that are eligible for the review include the following:

Behavioral Interventions: Sample interventions include functional behavioral assessments (FBAs) with accompanying support plans. These are data-driven interventions that are tailored to the needs of specific students. FBAs are sometimes delivered as part of wider prevention programs such as School-Wide Positive Behavior Interventions and Supports. Subcategories of behavioral interventions may focus on externalized behavioral concerns as well as internalized ones. It is appropriate to think of FBA as a replicable intervention technique, and reviewers need not focus on some of the idiosyncratic reinforcers and approaches to punishment.[3]

Academic Interventions: An overlap will often exist between behavioral and academic interventions because several educators argue that strong curricula can prevent behavioral concerns. Furthermore, the two will often intertwine, in that behavioral supports may entail reinforcement of academic skills and tasks. As long as sample criteria are met, any study of academic intervention impacts is eligible for review.

Social Skills Training: Some interventions have a strong focus on improving social skills among target children. The general goal of such training is to help recipients recognize (often subtle) social signals and react appropriately. These interventions are often delivered in group sessions but could be offered in individual therapy or as a component of a wider support plan.

Other Therapeutic Interventions and Consultation: Practically any form of school-based counseling directly provided to students, consulting services in which personnel train parents to provide home-based delivery of services, or teacher delivery of classroom-based programs is eligible for review.

Interventions that include the use of medication may be reviewed, but interventions that consist only of medication delivery are ineligible. On this point, many interventions will use simple, concrete elements, whereas others will apply multiple approaches. When a broad spectrum of features is delivered, the review will focus on aggregated findings. The impacts of subfeatures will be described in appendices if details are made available and related comparisons pass standards (probably limited to single-case designs).

Some interventions may encompass several grade levels (e.g., FBA). In these cases, intervention reports will disaggregate information and impacts by elementary (grades K–6), middle (grades 7 and 8), and high school (grades 9–12).[4]

Types of Research Studies to be Included

Studies must have become publically available in or after 1989. Only empirical studies using quantitative methods and inferential statistical analysis and that take the form of a randomized controlled trial (RCT) or use a regression-discontinuity (RD), quasi-experimental (QED), or a single-case experimental design (SCD) are eligible for this review.

[3] To clarify, FBA represents a replicable technique used to identify child-specific approaches for increasing or decreasing given behaviors. The approaches themselves often vary from one child to another because reinforcement and punishment are not universal; however, the techniques used to derive what works for a child will typically be replicable, so it may be helpful to aggregate FBA findings in intervention reports. Also note that we use the term *punishment* in the behavioral sense, meaning an environmental change that reduces a given behavior.

[4] Some schools do not follow this scheme (e.g., a middle school may encompass grades 6 through 8). When necessary, we will use study descriptions as the main determinant of which type of school and/or grade is in a sample.

Types of Outcomes to be Included

To be part of a review, a study must include at least one relevant child outcome that is intentionally targeted by the intervention and measured directly by administering an assessment to the child or conducting an observation of the child.

Outcomes to be included:[5]

- *External Behavior:* delinquent behaviors, substance abuse, active engagement, adaptive functioning, self-control, and operationalized assessments of behavior. These can be measured via teacher, parent, and self-rating scales.

- *Emotional/Internal Behavior:* indicators of thought disorders, depression, and overall adjustment/well-being. These can be measured via teacher, parent, and self-rating scales.

- *Social Outcomes:* degree of peer rejection/isolation, social-cognitive skills, prosocial behavior, and social interaction. These can be measured via teacher, parent, peer, and self-rating scales.

- *Reading Achievement/Literacy:* standardized assessments and progress measures (not grades).

- *Math Achievement:* standardized assessments and progress measures (not grades).

- *School Attendance:* including dropout rates, graduation rates, and attendance rates.

- *Other academic performance:* academic outcomes that combine reading and math achievement or measure achievement in other domains (not grades).

A study's rating will be based only on those measures (if any) that are not overaligned. Overalignment occurs when the outcome measure includes some of the same materials (such as books or passages) that are used in the intervention or is administered to the intervention group as part of the intervention. Outcome measures that are closely aligned or tailored to the intervention are likely to demonstrate larger effect sizes than those that are not. In these situations, the intervention group might have an unfair advantage over the comparison group, and the effect size would not be a fair indication of the intervention's effects. Outcome measures that are overaligned with the intervention will not be included in determining an intervention's rating for this review.

SCD studies must include at least three attempts to demonstrate an intervention effect at three different points in time or with three different phase repetitions. For a phase to qualify as an attempt to demonstrate an effect, the phase must have a minimum of five data points to meet evidence standards. Any phases based on fewer than three data points cannot be used to demonstrate the existence or lack of an effect (i.e., will not meet evidence standards). Rare circumstances might warrant a lower threshold of only one or two data points (e.g., students exhibiting extreme self-injurious behavior or students with selective mutism). In such cases, the PI will consult with content experts and document any departures from WWC standards.

[5] Outcomes related to IEP goal attainment will be included under the most relevant domain (e.g., behavioral, social, reading achievement/literacy, math achievement, or other academic performance).

The benefits of interventions for children with an ED are intended to be retained well past the end of the intervention. Indeed, measures taken several months or years after the intervention may provide strong evidence for an intervention's effectiveness. Thus, posttest and later measures are admissible. This review, however, prioritizes immediate posttest findings for developing intervention ratings and improvement indices because these findings are most prevalent, unless a study explicitly describes a follow-up outcome as the primary emphasis. The review will include additional follow-up findings, when available and appropriate, in appendices to the report. Follow-up effects may involve assessment from any point after the immediate intervention context, such as whether the effect generalizes across a full day to after several weeks.

The study must include at least one outcome measure with evidence of face validity, and for outcomes that are tests or scales, sufficient reliability assessed using the standards determined by the WWC, as described below.

The psychometric properties of outcomes will be described in intervention reports. If the reliability of each outcome measure is not specified in the research article, data from the test or scale's publisher or other sources may be used to establish the reliability of an outcome measure for the study population. If studies did not analyze the reliability of outcome measures using study data, and analyses by test publishers or other researchers did not include children with an ED, any other available evidence of the psychometric properties of the measure for the target population will be considered. A decision about the adequacy of the outcome measure will be made on a case-by-case basis in consultation with experts. Following is a list of WWC criteria for establishing minimal psychometric standards in QEDs and RCTs:

- Internal consistency: minimum of 0.60
- Temporal stability/test-retest reliability: minimum of 0.40
- Inter-rater reliability: minimum of 0.50 (percentage agreement, correlation, Kappa)

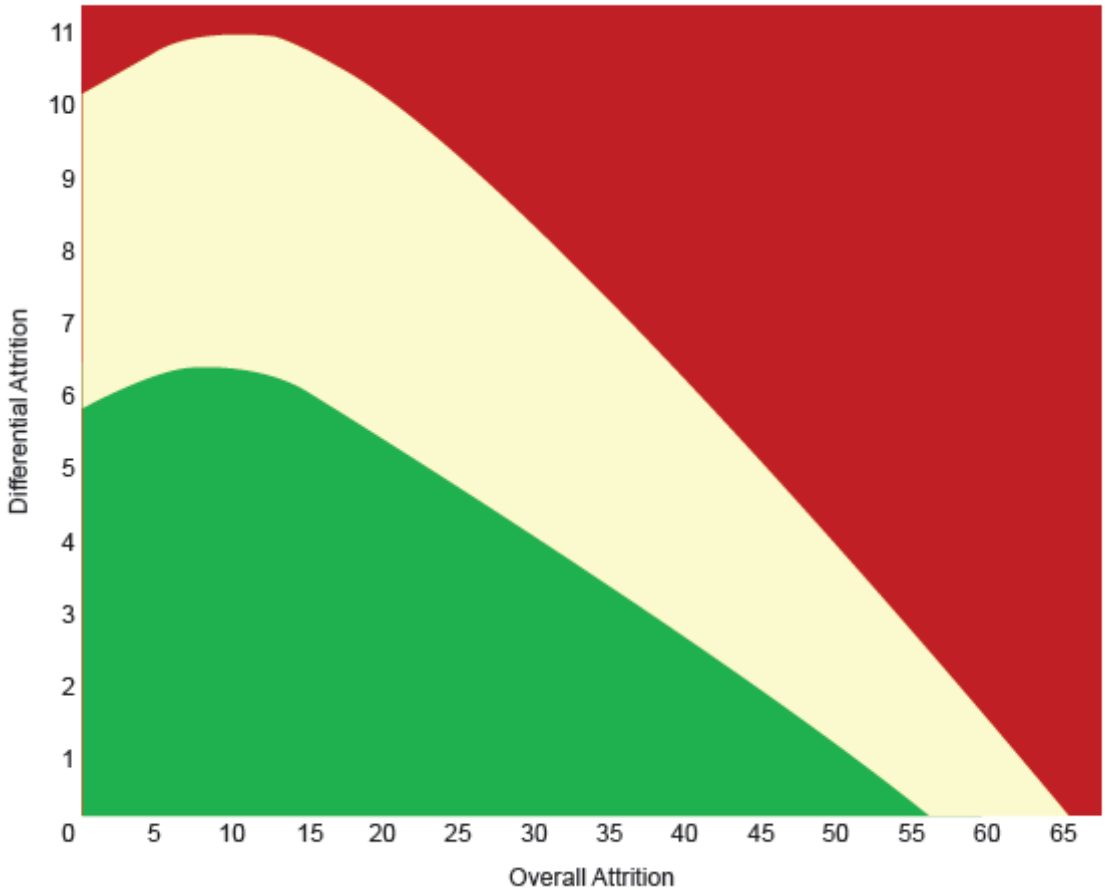
For SCD studies, the outcome variable must be measured systematically over time by more than one assessor. Inter-assessor agreement (commonly called *interobserver agreement*) must be documented on the basis of a statistical measure of assessor consistency; commonly used measures include percentage agreement (or proportional agreement) and Cohen's kappa coefficient. Minimum acceptable values of inter-assessor agreement range from 0.80 to 0.90 (on average) if measured by percentage agreement and at least 0.60 if measured by Cohen's kappa. Regardless of the statistic, inter-assessor agreement must be assessed for each case on each outcome variable. A study needs to collect inter-assessor agreement in all phases. It must also collect inter-assessor agreement on at least 20% of all sessions (total across phases) and on at least 20% of sessions within each condition (baseline and intervention.) If this standard is not met, the study *does not meet evidence standards*.

Attrition in RCTs and RDs

As described in the *WWC Procedures and Standards Handbook (version 2.0),* the WWC is concerned about overall and differential attrition from the intervention and comparison groups for RCTs, as both contribute to the potential bias of the estimated effect of an intervention. The attrition bias model developed by the WWC will be used in determining whether a study meets WWC evidence standards (see Appendix A of the *Handbook*).

When the combination of overall and differential attrition rates causes an RCT study to meet the liberal attrition standard (illustrated heuristically by the green and white areas on the diagram shown below), the attrition will be considered "low" and the level of bias acceptable. However, for RCTs with combinations of overall and differential attrition rates in the red area, the attrition will be considered "high" with potentially high levels of bias and, therefore, must demonstrate equivalence.

Many studies reviewed by the WWC are based on designs with multiple levels. Bias can be generated not only from the loss of clusters (such as schools), but also from sample members within the clusters (such as students), if those sample members attrit due to their treatment status. The attrition standard applies to both levels. To meet the standard, a study must first pass at the cluster level, using the designated attrition boundary. Second, the study must pass at the subcluster level, using the same attrition boundary, *with attrition based only on the clusters still in the sample.* That is, the denominator for the subcluster attrition calculation includes only sample members at schools or classrooms that remain in the study after cluster attrition.



Highest Level of Differential Attrition Allowable to Meet the Attrition Standard Under the Liberal Attrition Standard

Overall Attrition	Allowable Differential Attrition	Overall Attrition	Allowable Differential Attrition
0	10.0	34	7.4
1	10.1	35	7.2
2	10.2	36	7.0
3	10.3	37	6.7
4	10.4	38	6.5
5	10.5	39	6.3
6	10.7	40	6.0
7	10.8	41	5.8
8	10.9	42	5.6
9	10.9	43	5.3
10	10.9	44	5.1
11	10.9	45	4.9
12	10.9	46	4.6
13	10.8	47	4.4
14	10.8	48	4.2
15	10.7	49	3.9
16	10.6	50	3.7
17	10.5	51	3.5
18	10.3	52	3.2
19	10.2	53	3.0
20	10.0	54	2.8
21	9.9	55	2.6
22	9.7	56	2.3
23	9.5	57	2.1
24	9.4	58	1.9
25	9.2	59	1.6
26	9.0	60	1.4
27	8.8	61	1.1
28	8.6	62	0.9
29	8.4	63	0.7
30	8.2	64	0.5
31	8.0	65	0.3
32	7.8	66	0.0
33	7.6	67	-

Group Equivalence in RCTs/RDs with High Attrition and QEDs

If the study design is an RCT or RD with high levels of attrition or is a QED, the study must demonstrate baseline equivalence of the intervention and comparison groups for the analytic sample. The onus for demonstrating equivalence in these studies rests with the authors of the work. Sufficient reporting of pre-intervention data should be included in the study report (or obtained from the study authors) to allow the review team to draw conclusions about the equivalence of the intervention and comparison groups.

For this topic area, it is possible for a study to meet evidence standards in one (or more) domains and not in others. Thus, rules for establishing baseline equivalence should be applied *within each domain*. However, because this topic area focuses on meeting the needs of children with emotional and behavioral disorders, baseline equivalence *must be established within at least one* of the following domains or the study will not meet evidence standards: external behavior, emotional/internal behavior, or social outcomes. A study must establish baseline equivalence on one or more of these behavioral/social characteristics regardless of the outcome domains reported at posttest (unless it is an RCT without attrition concerns). Equating is done via the use of a pretest assessment of behavioral/social concerns, but matching on sample characteristics can be used in lieu of a baseline measure. Eligible sample characteristics for this purpose include the following:

(1) ED classification status (i.e., the study reports that there are comparable numbers of classified children in the treatment and control groups).
(2) Presence of psychiatric disorders is comparable across study groups.
(3) Children in the analytic intervention and control groups were disciplined by school staff and/or repeatedly referred for specialized consults to address behavioral concerns.

Analytic intervention and control samples are considered equivalent if the reported differences in their pre-intervention behavioral/social characteristics are less than or equal to one-quarter of the pooled standard deviation, regardless of statistical significance.[6] However, if differences are greater than 0.05 standard deviations, analyses must statistically control for the individual-level pre-intervention characteristic(s) on which the groups differ. *If baseline differences exceed 0.25 standard deviations for any of the behavioral/social pretests, the study will not meet evidence standards.*

Recall that our definition of at risk *does not* include students who are described as at risk simply because of geographic location, race/gender, socioeconomic status, or academic standing because these are not adequate indicators of prolonged behavioral concerns. Studies that equate samples on these characteristics alone will not pass standards for this review because they would not demonstrate that the sample is equivalent in terms of its likelihood of developing pervasive and persistent behavioral concerns.

After establishing baseline equivalence on behavioral/social characteristics, baseline equivalence must also be demonstrated within the reading, math and/or other academic performance domains, if an eligible reading, math, and/or other academic performance outcome was measured. Analytic intervention and control samples are considered equivalent *within each of these domains* if the reported differences in their pre-intervention characteristics are less than or equal to one-quarter of the pooled standard deviation, regardless of statistical significance. However, if

[6] Group mean differences in proportions or differences in the probability of the occurrence of an event are treated as dichotomous (or binary) outcomes and are measured with an odds ratio.

differences *within the domain* are greater than 0.05 standard deviations, analyses must statistically control for the individual-level pre-intervention characteristic(s) on which the groups differ. If baseline differences exceed 0.25 standard deviations for *any* of the listed outcomes within a domain, the study will not meet evidence standards within this domain. Thus, if reading outcomes are reported, there must be evidence the sample was equated on a reading measure *in addition* to equating groups on the presence of behavioral/social concerns (i.e., equating on a reading measure alone is insufficient because this leaves open the possibility that behavioral differences between groups are responsible for or are masking treatment effects). The same standard holds for math and other academic performance outcomes.

Summary

QEDs (or RCTs with high attrition) must provide evidence that analytic groups were equated on some indicator of behavioral/social characteristics (e.g., pretest and/or sample matching). Failure to provide this evidence will mean that the study will not pass standards, regardless of the reported outcome domain. Once groups are equated on behavioral/social characteristics, outcomes in these domains can pass standards with reservations, as long as proper covariates are included, as needed. Outcomes in reading, math, or other academic performance also require equating within that domain.

Some specific examples follow:

- If a study reports outcomes in external behavior, internal behavior, social constructs, and/or school attendance, a single pretest assessment of behavior/social characteristics (or associated sample characteristics) can be sufficient for establishing that groups are equated. If differences on any of the behavior/social pretest(s) are greater than 0.05 standard deviations (but less than 0.25 standard deviations), analyses must statistically control for the individual-level pre-intervention characteristic(s) on which the groups differ.

- If outcomes in multiple domains are included in the study, they should be analyzed separately. For instance, if we have one outcome in the reading achievement domain and one outcome in the external behavior domain, and the pretest for reading achievement needs adjustment and the pretest for external behavior does not, the reading achievement pretest would not need to be included in the impact model for external behavior to still meet standards with reservations. The reading achievement model, however, could not meet standards with reservations unless reading achievement is adjusted for in that model.

- Baseline equivalence must be established within at least one of the following domains: external behavior, emotional/internal behavior, or social outcomes (or associated sample characteristics). Thus, if a study reports one outcome in the external behavior domain and three outcomes in the math domain, and baseline differences on the external behavior outcome exceed 0.25 standard deviations, the whole study will not meet evidence standards.

- Within each domain, one measure can serve as the pretest for multiple outcomes, if not all outcomes were measured at pretest. If we have outcomes A and B in the reading domain for which the pretest for A needs adjustment and there was no pretest for outcome B, pretest A would need to be adjusted for in the impact model for outcomes A and B. However, if both outcomes A and B were measured at pretest, and the pretest for

A needs adjustment but the pretest for B does not, then pretest A would need to be adjusted for in the impact model for outcome A, but not for outcome B.

In addition, if there is evidence that the populations were drawn from very different settings, the PI may decide that the environments are too dissimilar to provide an adequate comparison. The PI will consider the following characteristics and settings when making this determination:

- Percentage of children who are classified as having an ED versus at risk for being classified in the treatment and control conditions. This may be a particular concern in QEDs dealing with prevention programs because at-risk students may be less impaired.
- Severity of sample members' needs, including the presence of comorbid conditions.
- Whether services were delivered in self-contained settings versus inclusive environments.
- Whether services were delivered outside of school (but still under the guidance of school-based staff).
- Age, gender, socioeconomic status, race/ethnicity, and/or English Language Learner status of sample members.

Statistical and Analytical Issues

RCT studies with low attrition do not need to use statistical controls in the analysis, although statistical adjustment for well-implemented RCTs is permissible and can help generate more precise effect-size estimates. For RCTs, the effect-size estimates will be adjusted for differences in pre-intervention characteristics at baseline (if available) using a difference-in-differences method if the authors did not adjust for the pretest (see the *WWC Procedures and Standards Handbook*). Beyond the pre-intervention characteristics required by the equivalence standard, statistical adjustment can be made for other measures in the analysis as well, although they are not required.

For the WWC review, the preference is to report on and calculate effect sizes for post-intervention means adjusted for the pre-intervention measure. If a study reports both unadjusted and adjusted post-intervention means, the WWC review will report the adjusted means and unadjusted standard deviations. If adjusted post-intervention means are not reported, they will be requested from the authors.

The statistical significance of group differences will be recalculated if (1) the study authors did not calculate statistical significance, (2) the study authors did not account for clustering when there was a mismatch between the unit of assignment and unit of analysis, or (3) the study authors did not account for multiple comparisons when appropriate. Otherwise, the review team will accept the calculations provided in the study.

When a misaligned analysis is reported (i.e., the unit of analysis is not the same as the unit of assignment) and the authors are not able to provide a corrected analysis, the statistical significance of the effect sizes computed by the WWC will incorporate an adjustment for clustering. The default intra-class correlations used for this review are 0.20 for cognitive, language, literacy, and math outcomes, and 0.10 for social-emotional development, behavior and attitudes, functional abilities, and motor development outcomes. For an explanation of the clustering correction, see the *WWC Procedures and Standards Handbook.*

When multiple comparisons are made (i.e., multiple outcome measures are assessed within an outcome domain in one study) and not accounted for by the authors, the WWC accounts for this multiplicity by adjusting the reported statistical significance of the effect using the Benjamini-Hochberg correction. See the *WWC Procedures and Standards Handbook* for the formulas the WWC uses to adjust for multiple comparisons.

All standards apply to overall findings as well as to analyses of subsamples.

Single-Case Research

The following criteria apply for single-case research:

- The independent variable (i.e., the intervention) must be systematically manipulated, with the researcher determining when and how the independent variable conditions change.
- The outcome variable must be measured systematically over time by more than one assessor, and the study needs to collect inter-assessor agreement in all phases and at least 20% of all sessions (total across phases) for a condition (e.g., baseline, intervention). Studies that collect inter-assessor agreement in all phases and at least 20% of all sessions (total across phases), but in which it is not clear whether the 20% by condition requirement is met, will be included in intervention reports with a footnote.
- The study must include at least three attempts to demonstrate an intervention effect at three different points in time or with three different phase repetitions.
- For a phase to qualify as an attempt to demonstrate an effect that Meets Evidence Standards, the phase must have a minimum of five data points.
- For a phase to qualify as an attempt to demonstrate an effect that Meets Evidence Standards with Reservations, the phase must have a minimum of three data points
- Exception: For the purposes of this review there may be occasions when fewer than three data points in a phase will not require the study to be rated as Not Meeting Standards. The following are exceptions:
    - Interventions for severe problem behavior such as aggression and self-injury for which extended initial baselines or reversal conditions pose serious ethical and procedural concerns.
    - Interventions on "zero baseline" behaviors when there is no logical reason to believe that further assessment would yield other than zero baseline performance. An example of such a zero baseline performance may be when a child is asked to provide a verbal response and consistently provides no response to the request because the child has selective mutism. In such cases, a multiple probe design could be used in order to alleviate potential "punishing" effects of repeated failure experiences.

LITERATURE SEARCH METHODOLOGY

The review team will conduct a keyword search to identify interventions with studies that may be eligible for review. Then the team will conduct focused intervention searches to ensure that all potentially eligible studies of the identified interventions are identified. Each type of search is described below.

1. Keyword Search

Primary Objective: To identify interventions with potentially eligible studies and assess the likely extent of studies on each intervention. The focus will be on breadth rather than depth. Subsequent searches will focus on the selected interventions and be designed to capture *all* potentially eligible studies, including any that the keyword search did not identify.

Search Strategy: The following keywords are meant to capture literature that falls within the scope of the protocol. Targeted outcomes and study design terms are included to focus the search on identifying literature that will support an intervention report. The keyword list is followed by a list of databases that are searched.

Keyword List

Target Ages and Setting:

School-aged OR
School-based OR
Self-contained OR
K–12 OR
Elementary school OR
Middle school OR
Junior high OR
High school OR
Special education OR
Individual Education Program OR
IEP

AND

Target Disability:

Emotional Disturb* OR
Seriously Emotionally Disturb* OR
Emotional and Behavior* Disorder* OR
EBD OR
Emotionally handicapped OR
Depress* OR
Personality disorder OR
Antisocial personality disorder OR
Bipolar disorder OR
Conduct disorder OR
Disruptive behavior disorder OR
Dysthymia OR
Externalized behavior* problem OR
Internalized behavior* problem OR
Obsessive compulsive disorder OR
Oppositional defiant disorder OR
Post traumatic stress disorder OR
Schizophrenia OR
Selective mutism OR
Self-injurious behavior OR
Tourette's syndrome OR
Anxiety disorder OR
Somatic disorder

AND

Interventions:

Intervention* OR
Curricul* OR
Program* OR
Strateg* OR
Instruct* OR
Teach* OR
Train* OR
Technique* OR
Therap* OR
Approach*OR
Functional Behavioral Assessment OR
Positive Behavioral Interventions OR
Functional Analysis OR
Behavioral Intervention Plans OR
Functional Behavioral Analysis OR
Token economies OR
Behavioral support OR
Behavioral intervention OR
Cognitive behavioral intervention OR
Group contingenc*OR
Self-management OR
Time-out OR
Response cost OR
Social Skills Training OR
Cognitive Therap* OR
Individualized Education Plan OR
Psychotherapy
Group therap* OR
Tertiary prevention OR
Response to intervention OR
Schedules of reinforcement OR
Positive reinforcement OR
Operant conditioning

AND

Study Design:

Control group OR
Comparison group OR
Matched groups OR
Treatment OR
Random* OR
Assignment OR
Baseline OR
Experiment OR
Evaluation OR
Impact OR
Effectiveness OR
Causal OR
Posttest OR
Pretest OR
Randomized Control Trial OR
RCT OR
Quasi-experimental Design OR
QED OR
Regression discontinuity design OR
Changing criterion design OR
Intrasubject replication design OR
Multiple baseline design OR
Multi-element design OR
Multi element design OR
Single case design OR
Single subject design OR
ABAB design OR
Alternating treatment OR
Simultaneous treatment OR
Meta-analysis OR
Meta analysis OR
Reversal design OR
Withdrawal design

The core list of electronic databases that are typically searched across topics includes the following:

a. ***ERIC.*** Funded by the U.S. Department of Education, ERIC is a nationwide information network that acquires, catalogs, summarizes, and provides access to education information from all sources. All Department of Education publications are included in its inventory.

b. ***PsycINFO.*** PsycINFO contains more than 1.8 million citations and summaries of journal articles, book chapters, books, dissertations, and technical reports, all in the field of psychology. Journal coverage, which dates back to the 1800s, includes international material selected from more than 1,700 periodicals in more than 30 languages. More than 60,000 records are added each year.

c. ***Campbell Collaboration.*** C2-SPECTR (Social, Psychological, Educational, and Criminological Trials Register) is a registry of more than 10,000 randomized and possibly randomized trials in education, social work and welfare, and criminal justice.

d. ***Dissertation Abstracts.*** As described by Dialog, *Dissertation Abstracts* is a definitive subject, title, and author guide to virtually every American dissertation accepted at an accredited institution since 1861. Selected master's theses have been included since 1962. In addition, since 1988, the database includes citations for dissertations from 50 British universities that have been collected by and filmed at the British Document Supply Centre. Beginning with DAI Volume 49, Number 2 (Spring 1988), citations and abstracts from Section C, Worldwide Dissertations (formerly European Dissertations) have been included in the file. Abstracts are included for doctoral records from July 1980 (*Dissertation Abstracts International*, Volume 41, Number 1) to the present. Abstracts are included for master's theses from spring 1988 (*Masters Abstracts*, Volume 26, Number 1) to the present.

e. ***Academic Search Premier.*** This multidisciplinary database provides full text for more than 4,500 journals, including full text for more than 3,700 peer-reviewed titles. PDF backfiles to 1975 or further are available for well over 100 journals, and searchable cited references are provided for more than 1,000 titles.

f. ***EconLit.*** EconLit, the American Economic Association's electronic database, is the world's foremost source of references to economics literature. The database contains more than 785,000 records from 1969 to the present. EconLit covers virtually every area related to economics.

g. ***Business Source Corporate.*** This source contains full text from nearly 3,000 quality business and economics magazines and journals (including full text of many articles only abstracted in other sources we search). Information in this database dates as far back as 1965.

h. ***SocINDEX with Full Text.*** SocINDEX with Full Text is the world's most comprehensive and highest-quality sociology research database. The database features more than 1,986,000 records with subject headings from a 19,600+ term sociological thesaurus designed by subject experts and expert lexicographers. SocINDEX with Full Text contains full text for 708 journals dating back to 1908. This database also includes full text for more than 780 books and monographs and full text for 9,333 conference papers.

i. ***EJS E-Journals.*** Electronic journals from EBSCO host® provide article-level access for thousands of electronic journals available through EBSCO's Electronic Journal Service (EJS). This resource covers journals to which Mathematica subscribes.

j. ***Education Research Complete***. Education Research Complete is the definitive online resource for education research. Topics covered include all levels of education, from early childhood to higher education, and all educational specialties, such as multilingual education, health education, and testing. Education Research Complete provides indexing and abstracts for more than 1,840 journals as well as full text for more than 950 journals, and it includes full text for more than 81 books and monographs and for numerous education-related conference papers.

k. ***WorldCat***. WorldCat is the world's largest network of library content and services and allows users to search simultaneously the catalogs of more than 10,000 libraries, containing more than 1.2 billion books, dissertations, articles, CDs, and other media.

l. ***Cochrane Central Register of Controlled Trials.*** The Cochrane Central Register of Controlled Trials is a bibliography of controlled trials identified by contributors to the Cochrane Collaboration and others; it is part of an international effort to hand-search the world's journals and create an unbiased source of data for systematic reviews.

m. ***Cochrane Database of Systematic Reviews.*** Cochrane Database of Systematic Reviews contains full-text articles, as well as protocols focusing on the effects of health care. Data are evidence-based medicine and often are combined statistically (with meta-analysis) to increase the power of the findings of numerous studies, each too small to produce reliable results individually.

n. ***Database of Abstracts of Reviews of Effects***. Database of Abstracts of Reviews of Effects (DARE) includes abstracts of published systematic reviews on the effects of health care from around the world that have been critically analyzed according to a high standard of criteria. This database provides access to quality reviews in subjects for which a Cochrane review may not yet exist.

o. ***Cochrane Methodology Register***. The Cochrane Methodology Register (CMR) is a bibliography of publications that reports on methods used in the conduct of controlled trials. It includes journal articles, books, and conference proceedings; these articles are taken from the MEDLINE database and from hand searches. The database contains studies of methods used in reviews and more general methodological studies that could be relevant to anyone preparing systematic reviews. CMR records contain the title of the article, information on where it was published (bibliographic details), and in some cases a summary of the article. CMR is produced by the UK Cochrane Centre on behalf of the Cochrane Methodology Review Group.

p. ***CINAHL with Full Text.*** The Cumulative Index to Nursing and Allied Health Literature (CINAHL) with Full Text is the world's most comprehensive source of full text for nursing and allied health journals, providing full text for more than 600 journals indexed in CINAHL. This authoritative file contains full text for many of the most used journals in the CINAHL index with no embargo. Full-text coverage dates back to 1981.

In addition to the keyword search, the review team seeks to identify other relevant studies through the following approaches:

a. Public submissions:
   1) Materials submitted via the WWC website
   2) Materials submitted directly to WWC staff

b. Solicitations made to key researchers by the review team

c. Checking websites summarizing research on programs for children and youth, prior reviews, and research syntheses (i.e., using the reference lists of prior reviews and research syntheses to make sure key studies have not been omitted)

d. Searches of the websites of all the developers of relevant interventions or practices for any research or implementation reports

e. Searches of the websites of more than 50 think tanks, research centers, and associations that conduct research in this topic area

References resulting from these searches will be screened and sorted by intervention.

2. Intervention Search

Primary Objective: To identify *all* effectiveness studies conducted for a specific intervention identified in the keyword search.

Search Strategy:

- Conduct standard library searches of the intervention name (e.g., *Functional Behavioral Analysis*).[7]
- Scan references to identify possible synonyms for the intervention in the literature (such as "functional behavioral assessment"). Conduct standard library searches of these terms.
- Once some potentially eligible studies are identified, request full text and review the reference lists to cross-check search results. Similarly, review relevant literature reviews. Revise search terms as needed.
- Identify seminal researchers associated with the intervention. Conduct full text searches of the researcher name combined with the intervention name.
- Identify seminal studies of the intervention and conduct searches of the associated citation.

All references resulting from these searches will be screened for eligibility.

[7] A standard library search consists of searching titles and abstracts in each of the databases described above.

REFERENCES

Assistance to states for the education of children with disabilities and preschool grants for children with disabilities, final rule; child with a disability. (2006). *Federal Register 71*(156; 14 August), 46756.

Bower, E. M. (1981). *Early identification of emotionally handicapped children in school* (3rd ed.). Springfield, IL: Charles C. Thomas.

Forness, S. R., & Knitzer, J. (1992). A new proposed definition and terminology to replace "serious emotional disturbance" in Individuals with Disabilities Education Act. *School Psychology Review*, *21*, 12–20.

Kauffman, J. M., & Landrum, T. J. (2009). *Characteristics of emotional and behavioral disorders of children and youth* (9th ed.). Upper Saddle River, NJ: Merrill Prentice Hall.