Council for Exceptional Children Standards for Evidence-Based Practices in Special Education

Council for Exceptional Children

The voice and vision of special education

CEC MISSION

The Council for Exceptional Children is an international community of professionals who are the voice and vision of special education. CEC's mission is to improve, through excellence and advocacy, the education and quality of life for children and youth with exceptionalities and to enhance engagement of their families.

CEC VISION

The Council for Exceptional Children is a premier education organization, internationally renowned for its expertise and leadership, working collaboratively with strategic partners to ensure that children and youth with exceptionalities are valued and full participating members of society. As a diverse and vibrant professional community, CEC is a trusted voice in shaping education practice and policy.

2900 Crystal Drive, Suite 1000
Arlington, VA 22202
Voice: 703/620-3660
TTY: 703/264-9446
Fax: 703/264-9494
www.cec.sped.org

Council for Exceptional Children
The voice and vision of special education

Council for Exceptional Children Standards
for Evidence-Based Practices in Special Education

Printed in the United States of America

Acknowledgements

This report was commissioned by the Council for Exceptional Children Board of Directors. A workgroup comprised of seven special education researchers developed, vetted, and piloted the new standards for determining evidence-based practices (EBPs) in special education. CEC's goal is that the standards will be applied to better understand the effectiveness of a range of practices for learners with disabilities.

CEC President Robin Brewer acknowledged CEC's appreciation to its expert members in the workgroup including Bryan Cook, Chair, Viriginia Buysse, Janette Klingner, Tim Landrum, Robin McWilliam, Melody Tankersley, and Dave Test.

COUNCIL FOR EXCEPTIONAL CHILDREN

Standards for Evidence-Based Practices in Special Education

This statement presents an approach for categorizing the evidence base of practices in special education. The quality indicators and the criteria for categorizing the evidence base of special education practices is intended for use by groups or individuals with advanced training and experience in educational research design and methods.

These quality indicators and criteria only apply to studies examining the effect of an operationally defined practice or program on student outcomes. For example, programs or practices that improve instructor or parent behaviors, even if those behaviors have been shown to improve student outcomes, do not fall within the purview of this approach. Moreover, reviews of practices should be specific to an outcome area and learner population. That is, reviews should set clear parameters on a targeted outcome (e.g., reading comprehension) and a targeted learner population (e.g., children with learning disabilities, preschoolers with developmental delays, adolescents with autism, K–3 struggling readers, K–12 students with disabilities). Reviews might also be specific to a setting (e.g., public schools, inclusive classes) or type of interventionist (e.g., paraprofessionals).

Studies need not be published in a peer-reviewed journal to be included in a review using these standards. However, studies must be publicly accessible.

The work of Gersten and colleagues (2005) and Horner and colleagues (2005) guided the development of these standards, which may be viewed as a refinement of their foundational and exceptional scholarship. In developing the standards, CEC's EBP Workgroup also drew from a number of other sources for categorizing the evidence base of practices (e.g., What Works Clearinghouse) and incorporated the feedback of 23 anonymous special education researchers who kindly participated in a Delphi study. The Council for Exceptional Children (CEC) is indebted to Gersten et al., Horner et al., and the Delphi study participants, without whom this work would not have been possible.

Research Designs

CEC's approach to categorizing the evidence base of practices in special education considers two research methods: group comparison research (e.g., randomized experiments, nonrandomized quasi-experiments, regression discontinuity designs) and single-subject research. The rationale is that causality can be reasonably inferred from these designs when they are well designed and conducted.

In **experimental group comparison designs**, participants are divided into two or more groups to test the effects of a specific treatment manipulated by the researcher. The standards consider group comparison studies in which researchers study treatment and comparison groups through random (in randomized controlled trials) and non-random (e.g., group quasi-experimental designs, including regression discontinuity designs) assignment.

Single-subject experimental designs use participants (individuals or groups) as their own control and collect repeated measures of dependent variables over time to test the effects of a practice manipulated by the researcher. The standards consider single-subject designs that systematically address common threats to validity and reasonably demonstrate experimental control. For example, appropriately designed and conducted ABAB/reversal, multiple-baseline, changing-criterion, and alternating-treatment designs are acceptable. AB (i.e., baseline-intervention) designs, for example, are not considered.

Although CEC recognizes the important role that correlational, qualitative, and other descriptive research designs play in informing the field of special education, the standards do not consider research using these designs because identifying evidence-based practices involves making causal determinations, and causality cannot be reasonably inferred from these designs.

Quality Indicators

The intent of identifying quality indicators essential for methodologically sound, trustworthy intervention studies in special education is not to prescribe all the desirable elements of an ideal study, but to enable special education researchers to determine which studies have the minimal methodological features to merit confidence in their findings. CEC's approach to classifying the evidence base of practices considers the number and effects of group comparison and single-subject studies determined to be of sound methodological quality. *Methodologically sound studies* must meet all the quality indicators specified for the relevant research design. Requiring studies to address all quality indicators in order to be classified as methodologically sound will necessarily limit the consideration of studies conducted before quality indicators were developed and emphasized in published studies. However, this conservative approach increases the likelihood that only the highest quality and most trustworthy studies are considered when classifying the evidence base of practices.

Quality indicators (see Table 1) may be rated as met when the study under review addresses the underlying intent. A study is considered to have addressed a quality indicator when reviewers agree that the methodological issue is addressed satisfactorily such that it does not represent a meaningful threat to the validity of study findings.

Whether a quality indicator is addressed in a study is commonly determined explicitly by what is reported in a research report. However, reviewers might sometimes need to use their informed judgment to determine whether a quality indicator has been met. In these cases—when reviewers can reasonably infer that the quality indicator is met on the basis of other, related information reported—they can decide that a study has met a quality indicator, even if the research report does not explicitly report addressing it.

Although social validity is not assessed with a distinct set of quality indicators, all studies must have socially important outcomes (see the first quality indicator under Outcome Measures/Dependent Variables). Further, magnitude of change in outcome variables must be socially valid for studies to be classified as having positive effects (see Classifying Effects of Studies).

Table 1. Quality Indicators

	Quality indicator	Notes
1.0. Context and setting. ***The study provides sufficient information regarding the critical features of the context or setting.***		
	1.1. The study describes critical features of the context or setting relevant to the review; for example, type of program or classroom, type of school (e.g., public, private, charter, preschool), curriculum, geographic location, community setting, socioeconomic status, physical layout.	B
2.0. Participants. ***The study provides sufficient information to identify the population of participants to which results may be generalized and to determine or confirm whether the participants demonstrated the disability or difficulty of focus.***		
	2.1. The study describes participant demographics relevant to the review (e.g., gender, age/grade, race/ethnicity, socioeconomic status, language status).	B
	2.2. The study describes disability or risk status of the participants (e.g., specific learning disability, autism spectrum disorder, behavior problem, at risk for reading failure) and method for determining status (e.g., identified by school using state IDEA criteria, teacher nomination, standardized intelligence test, curriculum-based measurement probes, rating scale).	B
3.0. Intervention agent. ***The study provides sufficient information regarding the critical features of the intervention agent.***		
	3.1. The study describes the role of the intervention agent (e.g., teacher, researcher, paraprofessional, parent, volunteer, peer tutor, sibling, technological device/computer) and, as relevant to the review, background variables (e.g., race/ethnicity, educational background/licensure).	B
	3.2. The study describes any specific training (e.g., amount of training, training to a criterion) or qualifications (e.g., professional credential) required to implement the intervention, and indicates that the interventionist has achieved them.	B
4.0. Description of practice. ***The study provides sufficient information regarding the critical features of the practice (intervention), such that the practice is clearly understood and can be reasonably replicated.***		
	4.1. The study describes detailed intervention procedures (e.g., intervention components, instructional behaviors, critical or active elements, manualized or scripted procedures, dosage) and intervention agents' actions (e.g., prompts, verbalizations, physical behaviors, proximity), or cites one or more accessible sources that provide this information.	B
	4.2. When relevant, the study describes materials (e.g., manipulatives, worksheets, timers, cues, toys), or cites one or more accessible sources providing this information.	B
5.0. Implementation fidelity. ***The practice is implemented with fidelity.***		
	5.1. The study assesses and reports implementation fidelity related to adherence using direct, reliable measures (e.g., observations using a checklist	B

	of critical elements of the practice).	
	5.2. The study assesses and reports implementation fidelity related to dosage or exposure using direct, reliable measures (e.g., observations or self-report of the duration, frequency, curriculum coverage of implementation).	B
	5.3. As appropriate, the study assesses and reports implementation fidelity (a) regularly throughout implementation of the intervention (e.g., beginning, middle, end of the intervention period), and (b) for each interventionist, each setting, and each participant or other unit of analysis. If either adherence or dosage is assessed and reported, this item applies to the type of fidelity assessed. If neither adherence nor dosage is assessed and reported, this item is not applicable.	B
6.0. Internal validity. ***The independent variable is under the control of experimenter. The study describes the services provided in control and comparison conditions and phases. The research design provides sufficient evidence that the independent variable causes change in the dependent variable or variables. Participants stayed with the study, so attrition is not a significant threat to internal validity.***		
	6.1. The researcher controls and systematically manipulates the independent variable.	B
	6.2. The study describes baseline (single-subject studies) or control/comparison (group comparison studies) conditions, such as the curriculum, instruction, and interventions (e.g., definition, duration, length, frequency, learner: instructor ratio).	B
	6.3. Control/comparison-condition or baseline-condition participants have no or extremely limited access to the treatment intervention.	B
	6.4. The study clearly describes assignment to groups, which involves participants (or classrooms, schools, or other unit of analysis) being assigned to groups in one of the following ways: (a) randomly; (b) nonrandomly, but the comparison groups are matched very closely to the intervention group (e.g., matched on prior test scores, demographics, a propensity score; see Song & Herman, 2010); (c) nonrandomly, but techniques are used to measure differences and, if meaningful differences are identified—for example, statistically significant difference, difference greater than 5% of a standard deviation (What Works Clearinghouse, 2011)—to statistically control for any differences between groups on relevant pretest scores or demographic characteristics (e.g., statistically adjust for confounding variable through techniques such as ANCOVA or propensity score analysis); or (d) nonrandomly on the basis of a reasonable cutoff point (regression discontinuity design).	G
	6.5. The design provides at least three demonstrations of experimental effects at three different times.	S
	6.6. For single-subject research designs with a baseline phase (alternating treatment designs do not require a baseline), all baseline phases include at least three data points (except when fewer are justified by study author due to	S

	reasons such as measuring severe or dangerous problem behaviors and zero baseline behaviors with no likelihood of improvement without intervention) and establish a pattern that predicts undesirable future performance (e.g., increasing trend in problem behavior, consistently infrequent exhibition of appropriate behavior, highly variable behavior).	
	6.7. The design controls for common threats to internal validity (e.g., ambiguous temporal precedence, history, maturation, diffusion) so plausible, alternative explanations for findings can be reasonably ruled out. Commonly accepted designs such as reversal (ABAB), multiple-baseline, changing criterion, and alternating treatment address this quality indicator when properly designed and executed, although other approaches can be accepted if study authors justify how they ruled out alternative explanations for findings or control for common threats to internal validity.	S
	6.8. Overall attrition is low across groups (e.g., < 30% in a 1-year study).	G
	6.9. Differential attrition (between groups) is low (e.g., ≤10%) or is controlled for by adjusting for noncompleters (e.g., conducting intent-to-treat analysis).	G
7.0. Outcome measures/dependent variables. ***Outcome measures are applied appropriately to gauge the effect of the practice on study outcomes. Outcome measures demonstrate adequate psychometrics.***		
	7.1. Outcomes are socially important (e.g., they constitute or are theoretically or empirically linked to improved quality of life, an important developmental/learning outcome, or both).	B
	7.2. The study clearly defines and describes measurement of the dependent variables.	B
	7.3. The study reports the effects of the intervention on all measures of the outcome targeted by the review (p levels and effect sizes or data from which effect sizes can be calculated for group comparison studies; graphed data for single-subject studies), not just those for which a positive effect is found.	B
	7.4. Frequency and timing of outcome measures are appropriate. For most single-subject studies, a minimum of three data points per phase is necessary if a given phase is to be considered as part of a possible demonstration of experimental effect (except when fewer are justified by study author due to reasons such as measuring severe or dangerous problem behaviors and zero baseline behaviors with no likelihood of improvement without intervention). For alternating treatment designs, at least four repetitions of the alternating sequence are required (e.g., ABABABAB; see Kratochwill et al., 2013).	B
	7.5. The study provides evidence of adequate internal reliability, interobserver reliability, test-retest reliability, or parallel-form reliability, as relevant (e.g., score reliability coefficient ≥ .80, interobserver agreement ≥ 80%, kappa ≥ 60%).	B
	7.6. The study provides adequate evidence of validity, such as content, construct, criterion (concurrent or predictive), or social validity.	G
8.0. Data Analysis. ***Data analysis is conducted appropriately. The study reports information on effect size.***		
	8.1. Data analysis techniques are appropriate for comparing change in	G

	performance of two or more groups (e.g., t tests, ANOVAs/MANOVAs, ANCOVAs/MANCOVAs, hierarchical linear modeling, structural equation modeling). If atypical procedures are used, the study provides a rationale justifying the data analysis techniques.	
	8.2. The study provides a single-subject graph clearly representing outcome data across all study phases for each unit of analysis (e.g., individual, classroom, other group of individuals) to enable determination of the effects of the practice. Regardless of whether the study report includes visual or other analyses of data, graphs depicting all relevant dependent variables targeted by the review should be clear enough for reviewers to draw basic conclusions about experimental control using traditional visual analysis techniques (i.e., analysis of mean, level, trend, overlap, consistency of data patterns across phases).	S
	8.3. The study reports one or more appropriate effect size statistic (e.g., Cohen's d, Hedge's G, Glass's Δ, □2) for all outcomes relevant to the review being conducted, even if the outcome is not statistically significant, or provides data from which appropriate effect sizes can be calculated.	G

Note. B = applies to both group comparison and single-subject research studies; G = indicator applies only to group comparison studies; S = indicator applies only to single-subject research studies; IDEA = Individuals with Disabilities Education Act.

Classifying the Evidence Base of Practices

The criteria for evidence-based classifications use the study as the unit of analysis. Studies are considered only when they (a) use either a group comparison (e.g., randomized experiments, nonrandomized quasi-experiments, regression discontinuity design) or single-subject research (e.g., reversal, multiple baseline, changing criterion, alternating treatment) design, and (b) are categorized as methodologically sound. *Methodologically sound* studies meet all of the quality indicators listed in Table 1 relevant to their research design. On the basis of the quantity, effects, and research design of methodologically sound studies reviewed, practices are classified in one of five categories: evidence-based practices, potentially evidence-based practices, mixed effects, insufficient evidence, or negative effects.

Classifying Effects of Studies

The criteria for categorizing the evidence base of practices in special education require that methodologically sound studies be classified as having positive, neutral or mixed, or negative effects.

Group Comparison Studies. For group comparison studies, review teams set their own effect size criteria for classifying studies as having positive, neutral or mixed, or negative effects a priori. Effect size criteria must be justified, based factors such as what constitutes a meaningful (i.e., socially valid) level of improvement in student performance for the target population on the outcome variable and empirical benchmarks related to the outcome measure (e.g., specialized, researcher-developed measures are associated with larger effects than standardized measures), intervention (e.g., higher effects are associated with teaching techniques than whole-school

interventions), and targeted recipients (e.g., interventions targeting individuals and small groups are associated with larger effects than those targeting whole classes and schools; see Lipsey et al., 2012).

To illustrate, when conducting a review of a teaching technique that targets individual learners and is typically assessed using researcher-developed outcome measures, the research team should set relatively high criteria for effect sizes. For example, using the What Works Clearinghouse (2011) effect size guidelines ($d \geq 0.25$ = positive effects, $d \leq -0.25$ = negative effects, with neutral/mixed effects indicated by $-0.25 < d < 0.25$) as a baseline, the review team might set effect size cutoff of $d \geq 0.40$ = positive effects and $d \leq -0.40$ = negative effects, with neutral or mixed effects indicated by $-0.40 < d < 0.40$. Demonstrating the social validity of outcome changes associated with an effect size of 0.40 for the target population would further justify using these cutoff points for the review. Any effect size statistic appropriate for representing differences between groups can be used (e.g., Cohen's *d*, Hedge's *G*, Glass's Δ, $\square^2$).

If a study measures multiple outcomes, reviewers should only consider effect sizes for outcomes relevant to their review. For example, for a review of the effect of a practice on reading fluency, if a study measured effects of the practice on both reading fluency and reading comprehension, only the former effects should be considered. If a study examines the effects of a practice on two or more measures of the outcome of focus (e.g., two measures of reading fluency), those effect sizes should be averaged to compute a mean effect size for the study.

***Single-Subject Studies*.** Single-subject research studies are classified as having positive, neutral or mixed, or negative effects on the basis of (a) the number and proportion of participants in a study for whom a functional relationship between the independent variable and the dependent variable was established and (b) the direction of the functional relationships. The presence of a functional relationship is established by reviewers' use of standard methods of visual analysis. This may include examining data across phases for changes in (a) level (mean scores within phases), (b) trend (slope of data within a phase), and (c) variability (range of scores around a level or trend line), as well as assessments of the (d) immediacy of any observed treatment effect and (e) overlap of data points across phases (for additional information on visual analysis, see Kazdin, 2010; Kratochwill et al., 2013). Note that authors of research reports being reviewed might describe the extent to which a functional relationship is established and the effects are meaningful, as well as the methods used to determine such. However, reviewers should conduct their own visual analysis of data and graphs included in the report to draw their own conclusions.

A single-subject study is considered to have *positive effects* when a functional relationship is established between the independent and dependent variables, resulting in a meaningful, therapeutic change in the targeted dependent variable for at least three-fourths (75%) of the cases (depicted by tiers on a graph) in a study. There should be a minimum of three total cases, and the data for none of the cases show evidence of a functional relationship between the independent and dependent variables that results in change in a nontherapeutic (harmful) direction. For example, an ABAB study with four participants from the population targeted by the review would be classified as having positive effects if meaningfully positive, functional relationships are established for three

participants, even if no functional relationship is established between the independent and dependent variables for the fourth participant. The magnitude of change in the dependent variable is considered meaningful or therapeutic when it is socially or practically important (e.g., as determined by social comparison, subjective evaluation, or other evidence that change in outcome was meaningful for participants).

A study is considered to have *negative effects* when a functional relationship is established between the independent and dependent variables, resulting in a nontherapeutic change (i.e., data intended to increase actually decrease or vice versa) in the targeted dependent variables for at least three-fourths (75%) of relevant cases (e.g., participants from the population targeted by the review) in a study. There should be a minimum of three cases.

A study is considered to have *neutral or mixed effects* when the criteria for neither positive nor negative effects are met. For example, a study would be considered to have neutral or mixed effects if a functional relationship were established between the independent and dependent variables with positive effects for two of four participants in an ABAB study, but no effects or negative effects were observed for the remaining participants.

If a study measures multiple outcomes, reviewers should consider effects only for outcomes relevant to their review. For example, for a review of the effect of a practice on reading fluency, if a study measured effects of the practice on both reading fluency and reading comprehension, only the former effects should be considered. If a study examines the effects of a practice on two or more measures of the outcome of focus (e.g., two measures of reading fluency), each dependent variable is treated as a separate case when determining the overall effect of the study. For example, in an ABAB study examining the impact of a practice on two different measures of reading fluency for three participants, meaningful, positive effects of the IV on the DV must occur for at least five of the six graphed outcomes (≥ 75%) for the study to be categorized as having positive effects. In addition, reviewers should only consider participants who are part of population targeted by their review. For example, for a review targeting learners categorized as having emotional or behavioral disorders (EBD), in an ABAB study involving three participants with EBD and two with learning disabilities, reviewers only consider (i.e., visually analyze) the outcomes for participants with EBD.

Evidence-Based Classifications

These standards establish criteria for five evidence-based classifications: evidence-based practices, potentially evidence-based practices, mixed effects, insufficient evidence, or negative effects (see Table 2).

Table 2. Evidence-Based Classifications

1. Evidence-based practice	
(a)	Must be supported by at least

	• two methodologically sound group comparison studies with random assignment to groups, positive effects, and at least 60 total participants across studies; • four methodologically sound group comparison studies with non-random assignment to groups, positive effects, and at least 120 total participants across studies; or • five methodologically sound single-subject studies with positive effects and at least 20 total participants across studies; **OR**
(b)	Meet at least 50% of criteria for two or more of the study designs described in (a). For example, the practice is supported by • one methodologically sound group comparison study with random assignment, positive effects, and at least 30 total participants, as well as three methodologically sound single-subject research studies with positive effects and at least 10 total participants; or • three methodologically sound single-subject studies with positive effects and at least 10 total participants, as well as two methodologically sound group comparison studies with non-random assignment, positive effects, and at least 60 total participants; **AND**
(c)	Include no methodologically sound studies conducted with negative effects and at least a 3:1 ratio of methodologically sound studies with positive effects to methodologically sound studies with neutral/mixed effects. For this item, CEC considers group experimental, non-randomly assigned group comparison, and single-subject design studies collectively.
2. Potentially evidence-based practice	
(a)	Must be supported by • one methodologically sound group comparison study with random assignment to groups and positive effects; • two or three methodologically sound group comparison studies with non-random assignment to groups; and positive effects; or • two to four methodologically sound single subject studies with positive effects; **OR**
(b)	Meet at least 50% of criteria for two or more of the study designs described in (a). For example, practice is supported by one methodologically sound single-subject study with positive effects and one methodologically sound non-randomly assigned group comparison study with positive effects; **AND**
(c)	Include no methodologically sound studies conducted with negative effects, and at least a 2:1 ratio of methodologically sound studies with positive effects to methodologically sound studies with neutral/mixed effects. For this item, CEC considers group experimental, non-randomly assigned group comparison, and single-subject design studies collectively.
3. Mixed evidence	
(a)	Must meet criterion (a) or (b) for evidence-based practice or potentially evidence-based practice (regarding number of methodologically sound studies with positive effects supporting the practice) **AND**
(b)	The ratio of methodologically sound studies with positive effects to methodologically sound studies with neutral/mixed effects is less than 2:1; **OR** one or more methodologically sound studies conducted with negative effects, as long as methodologically sound studies with negative effects do not outnumber methodologically sound studies with positive effects.
4. Insufficient evidence	
	Insufficient research exists to meet the criteria for any of the other evidence-based

	categories.
5. Negative effects	
(a)	Must include more than one methodologically sound study (of any acceptable design) conducted with negative effects, **AND**
(b)	The number of methodologically sound studies conducted with negative effects outnumbers the number of methodologically sound studies with positive effects.

References

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children, 71*, 149–164.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*, 165–179.

Kazdin, A. E. (2010). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York, NY: Oxford University Press.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from http://ies.ed.gov/ncee/pdf/wwc_scd.pdf

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education*, *34*, 26–38. http://dx.doi.org/10.1177/0741932512452794

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., … Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncser/pubs/20133000/pdf/20133000.pdf

Song, M., & Herman, R. (2010). Critical issues and common pitfalls in designing and conducting impact studies in education: Lessons learned from the What Works Clearinghouse (Phase I). *Educational Evaluation and Policy Analysis*, *32*, 351–371. http://dx.doi.org/10.3102/0162373710373389

What Works Clearinghouse. (2011). *Procedures and standards handbook* (Version 2.1). Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v2_1_standards_handbook.pdf