CASEL *Guide to Effective Social & Emotional Learning Programs*

Evidence Criteria

To be considered for inclusion in the CASEL *Guide to Effective Social and Emotional Learning Programs*, a program or approach must have an evaluation that meets each of four evidence criteria. These criteria involve (a) the type of research design used, (b) the setting in which the program was implemented, (c) the statistical findings, and (d) the types of outcomes demonstrated in the evaluation. The *Program Guide* evidence criteria have been informed by other systematic review frameworks (e.g., *Blueprints for Healthy Youth Development*, *NREPP*, and *What Works Clearinghouse*), and preeminent standards for evaluation methodology (Flay et al., 2005; Gottfredson et al., 2015; Shadish, Cook, & Campbell, 2002). These criteria were developed by CASEL research staff and consultants, and an expert advisory panel reviewed and adjusted these criteria in November 2016[1] and August 2017.[2]

SELect programs

1. *The evaluation must use a pre-post randomized control trial (RCT) or pre-post quasi-experimental (QE) design that includes a comparison group that did not participate in the program.*

   An evaluation is considered to be an RCT when participants were assigned to conditions (i.e., intervention and comparison) on a random basis. A study is considered a QE design when group assignment was conducted through a process that was not random or when randomization was broken for any reason.

2. *The evaluation must assess the effects of a program that is implemented at the universal level, during the regular school day, with students who are within a grade range that spans from preschool through 12th.*

   At present, the focus of the *Program Guide* is on universal programs that are designed to be implemented with all students in a school. We do not include evaluations of programs that were conducted with students who were preselected based on a history or risk of a behavioral problem (e.g., students referred by staff for participation in the program based on history of oppositional behavior), nor evaluations conducted in a non-traditional school setting.[3] We make an exception for evaluations that preselected students based on a history or risk of academic failure and include these for review.

   Some programs are designed to be used at multiple levels of a school, including those levels where students can utilize more intensive supports (e.g., Tier II and treatment settings) and we

[1] Panel members included: Drs. Roisin P. Corcoran, Joseph A. Durlak, Linda Dusenbury, & Mark T. Greenberg.
[2] Dr. Kimberly Schonert-Reichl was added to the advisory panel.
[3] Non-traditional school settings are settings wherein students are referred or are enrolled based on students' level of functioning (e.g., high, low).

believe that these programs are beneficial. However, given the focus of the *Program Guide*, an evaluation is included only if it assessed the effects of a program when implemented at the universal level during the regular school day.

3. *The evaluation must report statistically significant main effects (at the $p < .05$ probability level or beyond) between the intervention and comparison group adjusting for outcome pretest and using clear and appropriate analytic methods.*

   Statistically adjusting for outcome pretest is required to minimize statistical bias when estimating effects of a program and to increase confidence in the accuracy of effects. According to the *What Works Clearinghouse* of the Institute for Educational Sciences (WWC), adjusting for important covariates such as outcome pretest is beneficial for both RCTs and QE designs as this provides greater statistical precision of parameter estimates, thus providing a more reliable test of program effectiveness (WWC, 2016a). Adjusting for pretest is particularly important for QE designs as these covariates might confound the effects of the intervention (WWC, 2016a). Even when no pretest differences are evident, adjusting for them will control for chance variations and improve the precision of the impact estimates (Flay et al., 2005).

   Additional analyses are required for RCTs and QE designs that utilize change score analyses that do not adjust for pretest. Specifically, additional analyses must be conducted that assess the difference in posttest means while statistically adjusting for outcome pretest. The rationale for this is that these two types of analysis answer different research questions. While change score analysis can be used to show that the intervention group changed at a different rate on substantive student outcomes, it is not a sufficient test of whether students who participated in the program outperform on those outcomes at posttest *if baseline levels are held constant* (Knapp & Schafer, 2009). In other words, change score analysis does not adjust for the impact of pretest on outcome posttest, potentially over- or underestimating the group mean difference (Dimitrov & Rumrill, 2003; WWC, 2016b).

   There is one exception made to the *Program Guide's* requirement that pretest scores be included in outcome analyses. This exception only applies to RCTs that have (a) an adequate number of units per group (i.e., approximately 30) at the unit of assignment and (b) do not have severe rates of attrition. Even when faithfully executed, RCTs with differential attrition across groups have reduced group equivalence which can make study results ambiguous (Gottfredson et al., 2015). The *What Works Clearinghouse Procedures and Standards Handbook Version 3.0* (WWC, 2016a) was used as a model for determining acceptable vs. unacceptable levels of attrition. Specifically, studies without pretest are only considered for inclusion if group assignment was randomized (RCTs) and the ratio of differential vs. overall attrition fell within the "acceptable range" for likelihood of bias (see Figure III.2 on p. 12 of the WWC 2016 handbook).

   In addition, evaluations must not exclude sample units from analyses based on implementation factors when estimating intervention effects. A purpose of the *Program Guide* is to identify programs that have potential for broad dissemination, which includes consideration of how well a program can be scaled and implemented in real-world settings. Since low implementation quality

could occur when aspects of a program make it less appropriate or less feasible to implement, excluding low implementers from an analytic sample can weaken the generalizability of findings to natural school settings.

4. *Positive effects must be found with students on outcomes in at least one of the following behavioral student outcome domains: (1) academic performance, (2) positive social behavior, (3) problem behaviors, or (4) emotional distress. Furthermore, the preponderance of evidence from an evaluation must indicate positive effects on these measures for students in the intervention group and must not favor the comparison group that did not participate in the intervention.*

Behavioral Student Outcomes

Behavioral student outcomes are assessments of student's activity or performance over time and include self-, parent-, or teacher ratings, or the results of classroom or school observations conducted by outside parties.[4] There are four categories of qualifying student behavioral outcomes, including (1) improved academic performance, (2) improved positive social behavior, (3) reduced problem behaviors, and (4) reduced emotional distress.

Improved Academic Performance: Significant impact on academic performance (e.g., GPA, graduation rates, standardized test scores) that favor the intervention group.

Improved Positive Social Behavior: Significant impact on measures of positive social behavior (e.g., works well with others, positive peer relations, assertiveness, conflict resolution) that favor the intervention group.

Reduced Problem Behaviors: Significant impact on problem behaviors (e.g., aggressive or disruptive behavior, delinquency, substance use, dropout) that favor the intervention group.

Reduced Emotional Distress: Significant impact on emotional distress (e.g., depressive symptoms, anxiety, social withdrawal).

Additional Student Outcomes

Additional student outcomes are coded during the review process that are not considered to be behavioral student outcomes. These outcomes include measures of students' perception of their social or emotional skills, values, or intentions, and student attitudes about school, such as quality of the classroom climate or relationships between teachers and students. Findings on any additional student outcomes do not qualify a program for designation as "SELect", but they are noted for programs that otherwise meet these evidence criteria. Furthermore, if positive effects

[4] Examples of outcomes that would be coded as "behavioral" include: student's self-report of the extent to which they were engaged in prosocial behavior during a specified span of time; teacher's, family member's, or peer's report of the extent to which a student engaged in prosocial behavior during a specified span of time; a trained observer's systematic observation of a classroom and subsequent ratings of the extent to which students behaved prosocially during that interval.

are found with students on additional student outcomes, a program or approach could be considered for "Promising" designation (see Promising program section below).

Improved SEL Skills & Attitudes: Significant impact on students' SEL skills, attitudes, beliefs, mindsets, values, and/or intentions (e.g., perception of goal setting skills, intention to intervene if another student was being discriminated against or bullied, perception of student-teacher relationships and/or student-student relationships, feelings of connectedness to school) that favor the intervention group.

Teacher Outcomes

Teacher outcomes include measures of teachers' use of instructional practices that foster student social and emotional development and/or help create and establish a supportive learning environment. Demonstration of a teacher outcome on its own does not qualify an evaluation as "SELect", but they are noted for programs that otherwise meet these evidence criteria. Furthermore, if positive effects are found for improved teachers outcomes a program or approach could be considered for "Promising" designation (see Promising program section below).

Improved Teaching Practices: Significant impact on teaching practices that promote social and emotional development (e.g., teacher self-report of using emotional support techniques with students in their classroom; student self-report of their teacher's use of instructional practices that promote metacognition; ratings made by a trained observer on a systematic observation protocol indicating positive teacher-student interactions) that favor the intervention group.

Outcomes Favoring the Comparison Group

Outcomes are considered to favor the comparison group when the intervention group (a) demonstrates significantly lower scores on a qualifying positive student behavioral outcome relative to the comparison group or (b) demonstrates significantly higher scores on a qualifying negative student behavioral outcome relative to the comparison group. An evaluation cannot qualify a program for SELect designation if it demonstrates an effect that favors the comparison group on a qualifying student behavioral outcome. However, if an evaluation indicates effects that favor the comparison group but the preponderance of that evaluation's effects favor the intervention group by at least a 2:1 ratio (i.e., at least two effects favoring the intervention group for every one effect favoring the comparison group), that program could be designated as Promising.

Promising programs

The Promising designation is used when programs meet CASEL's program design criteria but do not meet the full evidence criteria. A program could be designated as Promising:

1. If a qualifying evaluation shows a positive effect favoring the intervention group on a nonbehavioral outcome such as attitudes (e.g., feelings of connectedness to school) or student

perception of their own social-emotional skills (e.g., emotion recognition or decision-making ability) but the evaluation does not also demonstrate a behavioral student outcome.

2. If a qualifying evaluation demonstrates an outcome for improved teaching practices but does not also demonstrate a behavioral student outcome.

3. If a qualifying evaluation indicates that the intervention group demonstrates significantly greater rates of positive change, but the analyses did not adjust for pretest scores, they could be considered for Promising if the evaluation demonstrates that groups were equivalent at baseline.

4. If a qualifying evaluation indicates effects that favor the comparison group on behavioral student outcomes, but the preponderance of that evaluation's effects favor the intervention group by at least a 2:1 ratio (i.e., at least two effects favoring the intervention group for every one effect favoring the comparison group).

Promising programs are eligible to become SELect once an additional evaluation with an independent sample meets all of CASEL's evidence criteria for SELect programs.

If an evaluation does not have effects favoring the intervention group by at least a 2:1 ratio, that evaluation is excluded from the *Program Guide* review process. In the instances where an evaluation was excluded from the review process, an additional evaluation would be needed to qualify that program for inclusion. The additional evaluation would need to meet the evidence criteria for this *Guide*, and would need to be conducted with an independent sample that demonstrates effects favoring the intervention group by at least a 2:1 ratio. This additional evaluation would allow the program to be considered for Promising designation.

References

Dimitrov, D. M., & Rumrill Jr, P. D. (2003). Pretest-posttest designs and measurement of change. *Work: Journal Of Prevention, Assessment & Rehabilitation*, *20*(2), 159-165.

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

Gottfredson, D., Cook, T., Gardner, F. M., Gorman-Smith, D., Howe, G., Sandler, I., & Zafft, K. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science*, 1-34.

Knapp, T.R., & Schafer, W.D. (2009). From gain score t to ANCOVA F (and vice versa). *Practical Assessment, Research & Evaluation, 14(6).*

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

What Works Clearinghouse. (2016a). *What Works Clearinghouse: Procedures and standards handbook (Version 3.0)*. Retrieved from http://ies.ed.gov/ncee/wwc/Handbooks.

What Works Clearinghouse (2016b). *Reviewer guidance for use with the procedures and standards handbook (Version 3.0).* Retrieved from http://ies.ed.gov/ncee/wwc/Document/260.