**What Works Clearinghouse Extent of Evidence Categorization**

The Extent of Evidence Categorization was developed to tell readers how much evidence was used to determine the intervention rating, focusing on the number and sizes of studies. This scheme has two categories: small and moderate/large.

- The extent of evidence is moderate/large:
  - The domain includes more than one study; AND
  - The domain includes more than one school; AND
  - The domain findings are based on a total sample size of at least 350 students OR, assuming 25 students in a class, a total of at least 14 classrooms across studies.

- The extent of evidence is small:
  - The domain includes only one study; OR
  - The domain includes only one school; OR
  - The domain findings are based on a total sample size of less than 350 students AND, assuming 25 students in a class, a total of less than 14 classrooms across studies.

Each intervention domain receives its own categorization. For example, each of the three domains in character education—behavior; knowledge, attitudes, and values, and academic achievement—receives a separate categorization.

Example:
Intervention Do Good, a character education intervention, had three studies that met WWC standards and were included in the review. All three studies reported on academic achievement. There were a total of 6 schools across the three studies. The first study reported testing on 150 students, the second study 125 students, and the third study reported testing 4 classes with 15 students in each class. The extent of evidence on academic achievement for the Do Good intervention is considered "moderate/large" – it met the condition for both the number of studies and the number of schools, and although the total number of students is less than 350 (150+125+(4*15)=335), the number of classes exceeded 14 (150/25+125/25+4=15).

A "small" extent of evidence indicates that the amount of the evidence is low. There is currently no consensus in the field on what constitutes a "large" or "small" study or database. Therefore, the WWC set the conditions above based on the following rationale:

- When there is only one study, there is the possibility that some characteristics of the study—the outcome instruments, the timing of the intervention, etc.—might have affected the findings. When there are multiple studies, especially if they differ, provide some assurance that the effects can be attributed to the intervention, and not some features of the particular place where the intervention was studied. Therefore, the WWC determined that the extent of evidence is small when the findings are based on only one setting.
- Similarly, when there is only one school, there is a possibility that some characteristics of the school—the principal, student demographics, etc.—might have affected the findings or are

intertwined or confounded with the findings. Therefore, the WWC determined that the extent of evidence is small when the findings are based on only a single school.

- The sample size of 350 was derived from the following assumptions:
  - a balanced sampling design that randomizes at the student level,
  - a minimum detectable effect size of 0.3,
  - the power of the test at 0.8,
  - a two-tailed test with an alpha of 0.05, and
  - the outcome was not adjusted by an appropriate pretest covariate.

The Extent of Evidence Categorization provided in recent reports, and described here, signals WWC's intent to eventually provide a rating scheme on the external validity, or the generalizability, of the findings, for which the extent of evidence is only one of the dimensions. The Extent of Evidence Categorization, in its current form, is not a rating on external validity; instead, it serves as an indicator that cautions readers when findings are drawn from studies with small size samples, a small number of school settings, or a single study.