IES Institute of Education Sciences

APPENDIXES
January 2021

Regional Educational Laboratory Southeast
At Florida State University

Effects of an Inquiry-Oriented Curriculum and Professional Development Program on Grade 7 Students' Understanding of Statistics and on Statistics Instruction

See https://go.usa.gov/xAXRK for the full report.

Appendix A. About the study

This cluster-randomized controlled trial compared the Supporting Teacher Enactment of the Probability and Statistics Standards (STEPSS) program—a 20-day curriculum unit designed to support teaching and learning of the state curriculum standards for statistics and probability in grade 7, along with four days of professional development for teachers—with practice-as-usual probability and statistics instruction and teacher professional development. The two primary outcomes of interest were student understanding of statistics and classroom instruction in statistics.

Setting

This study was conducted in 40 middle schools in Broward County Public Schools (BCPS), a large, urban school district in Florida. BCPS is the second largest school district in Florida and the sixth largest in the United States. It serves more than 250,000 K–12 students who speak more than 150 languages.

At the time of random assignment in 2017/18, the percentage of economically disadvantaged students in these 40 schools ranged from 18 percent to 96 percent, with a mean of 71 percent. In the same school year the percentage of students scoring proficient on the grade 7 state mathematics state assessment in the 40 schools ranged from 6 percent to 87 percent, with a mean of 47 percent. Grade 7 students included in the study sample, after accounting for attrition, reflect the following demographic characteristics: 34 percent Hispanic, 32 percent White, 28 percent Black, 4 percent Asian, and 46 percent male.

In Florida there are two grade 7 mathematics courses: regular and advanced. The course description for regular grade 7 mathematics contains fewer content standards than the course description for advanced grade 7 mathematics, which includes some content standards for grade 8 mathematics, but the standards from the probability and statistics domain in the two course descriptions are identical.

Description of the treatment

The STEPSS program involves two complementary components: four days of teacher professional development workshops in statistics, conducted in two-day intervals, and implementation of 12 lessons designed to teach the state curriculum standards in the statistics and probability domain in grade 7. The study team relied on experts at the American Statistical Association (ASA) to identify the 12 lessons for the study and to design the professional development. The ASA based the teacher workshops on its existing Meeting-within-a-Meeting program

(http://www.amstat.org/education/mwm), the goal of which is to increase teachers' understanding of statistics and provide them with resources and insight into effectively teaching statistics. Experienced trainers who were vetted and recommended by ASA officials led the STEPSS program workshops. Two of the workshop days occurred in the summer, as in the ASA's existing model, and two days occurred during the school year.

The content of the workshops and associated lesson plans in the STEPSS program focused on the probability and statistics standards that grade 7 students are expected to meet by the end of the school year. These standards set ambitious goals for students to develop a foundational understanding of statistics by including such topics as learning that statistics is the study of variability, the importance of representative samples, and some of the fundamental concepts supporting an understanding of how probability supports statistical inference. Most of the workshop time consisted of teachers participating in the same 12 lessons that they would to teach to students, thereby maximizing the coherence between the professional development and curriculum components of the treatment. The workshop leaders allocated the remaining time to a facilitated session for teachers to practice teaching the lessons in a microteaching lesson-study format (Fernández, 2005, 2010; Zhou & Xu, 2017).

Teachers in schools assigned to the STEPSS program received a set of 12 lesson plans, designed as a replacement unit to support teaching and learning of probability and statistics in regular and advanced grade 7 mathematics courses (table A1). The curriculum standards for the replacement unit were the same for both courses. The lesson plans centered on cognitively complex tasks that involve students in statistical investigation and inquiry. All of the lessons are published by the American Statistical Association and available in *Bridging the Gap Between Common Core State Standards and Teaching Statistics* (Hopfensperger et al., 2012) or in Statistics Education Web (http://www.amstat.org/education/STEW), an online database of peer-reviewed lesson plans. The instructional design model used in each lesson of the 12-lesson unit are consistent with the inquiry-oriented model recommended by the American Statistical Association in *Guidelines for Assessment and Instruction in Statistics Education Report: A PreK–12 Curriculum Framework* (Franklin et al., 2007). The four parts in the model involve students in formulating a question that can be answered by data, designing and implementing a data collection plan, analyzing the data using graphical and numerical methods, and interpreting the analyses in the context of the original question.

Table A1. Lessons and activities included in the Supporting Teacher Enactment of the Probability and Statistics Standards program

Lesson or activity	Primary source	Estimated number of classroom days[a]	Associated topic in the Mathematics Florida Standards
Formulating a Statistical Question	Bridging the Gap[b]	1	6.SP.1.1
Who Has the Longest First Name?	Bridging the Gap	2	6.SP.1.1 6.SP.1.2 6.SP.1.3
How Long Are Our Shoes?	Bridging the Gap	2	6.SP.1.1 6.SP.1.2 6.SP.1.3
How Large Are Families Today?	Statistics Education Web[c]	2	6.SP.1.1 6.SP.1.2 6.SP.1.3 6.SP.2.4 6.SP.2.5 7.SP.2.4
How Long Are the Words in the Gettysburg Address?	Statistics Education Web	2	7.SP.1.1 7.SP.1.2
How Far Can You Jump?	Bridging the Gap	2	7.SP.2.3 7.SP.2.4
How Fast Can You Sort Cards?	Bridging the Gap	2	7.SP.2.3 7.SP.2.4
Is There a Meaningful Difference Between the Number of Times People Can Write Their Name with Their Dominant Hand Versus Their Non-Dominant Hand?	Statistics Education Web	1	7.SP.2.3 7.SP.2.4
How Likely Is it?	Bridging the Gap	1	7.SP.3.5
What is the Chance of Seeing an Elephant at the Zoo?	Bridging the Gap	2	7.SP.3.5 7.SP.3.8
What Do Frogs Eat?	Bridging the Gap	1	7.SP.3.5 7.SP.3.7 7.SP.3.8
How Many Spins to Win the Prize?	Bridging the Gap	2	7.SP.3.6

a. Assumes a 45- or 50-minute class period.
b. Hopfensperger et al., 2012 (https://www.statisticsteacher.org/statistics-teacher-publications/).
c. https://www.amstat.org/ASA/Education/STEW/home.aspx.
Source: Authors' compilation.

Cost of the Supporting Teacher Enactment of the Probability and Statistics Standards program

The primary costs, specific to the STEPSS program, included the workshop leader fees, teacher pay for attendance in workshops, substitute teacher costs, travel costs, and cost of materials to implement the lessons. Other costs, such as the administrative costs for planning, teacher salaries and benefits, and building costs (electricity, custodial services, and rent), are not included in these figures because this cost analysis examines the incremental cost of the STEPSS program beyond the cost of practice as usual. The costs of the treatment are described with the goal of calculating the unit cost per teacher and student. Four days of workshops were offered for each teacher. Each pair of workshop leaders can teach a cohort of up to 30 teachers at a time, so the costs can be scaled up or down in units of 30 teachers.

Workshop leaders were paid $1,600 per day, which covered time for planning and other preparations, delivery of the workshops, and travel time. Two workshop leaders co-teach each workshop day. The cost of workshop leader fees for each cohort of up to 30 teachers was $1,600 per day x 4 days of workshops x 2 workshop leaders, for a total of $12,800.

The study team compensated teachers from schools assigned to the STEPSS program for their participation in the two days of summer workshops, which occurred outside their contracted time. In addition, teachers who completed the training received district credit for the corresponding number of hours of in-service training toward continuation of their teaching credentials.

Teachers were paid $150 per day for their attendance in the two days of summer workshops. Substitute teachers were paid $120 per day for the two days when teachers attended workshops during the school year. Assuming a cohort of 30 teachers, the expected cost for teachers and substitutes is 30 teachers x [($150 per day for summer days x 2 days) + ($120 per day for substitute teachers on school days x 2 school days)], for a total of $16,200.

Workshop leaders traveled to the school sites by plane or by car. They stayed in hotels for each visit, which involved two nights of lodging for each two-day session. A rate of $36 per day was provided for food costs. Total travel cost was approximately $900 per workshop leader for each two-day session. Travel costs per cohort amounted to $900 per two-day session x 2 two-day sessions x 2 people, for a total of $3,600.

At the time of this writing, the American Statistical Association provided public access to the lesson plans, free of charge. They are available in electronic format and printable. Printed copies were provided for teachers in this study at a cost of approximately $8 per teacher. At a rate of $8 per teacher, a set of these materials costs $240 per cohort of 30 teachers.

To make it easier for the teachers to implement the lessons, consumable booklets were printed for students. The cost of printing the booklets was approximately $1 per student. The intent-to-treat sample had an average of 92 students per teacher. The per cohort cost is $1 per student x 92 students x 30 teachers, for a total of $2,760.

Some lesson plans involve specialized materials, such as masking tape, rulers, animal crackers, and so on. These materials were provided for teachers at an approximate cost of $30 per teacher for a per cohort total of $900.

Using these cost estimates and the number of teachers per cohort (n_2 = 30) and average number of students of those teachers (n_1 = 92), the cost of the STEPSS program is approximately $1,208.88 per teacher or $13.14 per student.

Description of practice as usual

Teachers of regular and advanced grade 7 mathematics in practice-as-usual schools made no changes in how they taught statistics or participated in statistics-related professional learning opportunities. The recommended pacing guide from the BCPS curriculum department allocated 24 days for the unit on probability and statistics in the regular grade 7 mathematics course and 19 days in the advanced grade 7 mathematics course. The district-adopted textbook was Houghton Mifflin Harcourt's *GO Math!* (Burger et al., 2016). The number of days that teachers actually allocated for the probability and statistics unit is not known, and the source of the lesson plans is known only for the days when the study team observed probability or statistics instruction (see table A4 later in the appendix). The extent to which teachers in practice-as-usual schools participated in professional development opportunities in probability or statistics during the treatment period is also unknown, but curriculum leaders from the school districts report few or no professional development opportunities for teachers in this content area.

Eligibility

The study team met with BCPS leaders in March 2018 to begin identifying eligible schools. To be eligible to participate in the study, middle schools needed to have at least 30 students enrolled in a regular or advanced grade 7 mathematics course and two or more teachers responsible for teaching those courses.

From a list of all the schools offering regular or advanced grade 7 mathematics during the 2017/18 school year, several schools were identified as elementary schools, and those schools were deemed ineligible for the study.

BCPS leaders excluded a district virtual school because replacing the curriculum there was not feasible. All the remaining schools with at least 30 students enrolled in these courses were selected for further review. This resulted in a list of 47 schools, some of which served students in grades 6–8 and some of which served students in grades K–8 or 6–12. Seven of the 47 schools were excluded because they had only one teacher of regular or advanced grade 7 mathematics, in addition to having other unique characteristics. The remaining 40 schools met the initial eligibility criteria for the study. The study included every regular BCPS middle school with at least 30 students enrolled in regular or advanced grade 7 mathematics and two or more grade 7 mathematics teachers.

Random assignment

The study team used the SPSS Statistics 21 random sample of cases option to select a random sample of 20 schools from the list of 40 schools to participate in the STEPSS program and set the remaining 20 schools to participate in practice as usual. One class was randomly selected from the set of classes taught by each grade 7 mathematics teacher to be observed by the study team. Some teachers taught a single grade 7 mathematics class, so the study team selected that class for observation. For teachers who taught more than one class, the study team selected one class at random from among the classes that were offered in each class period in each school until all the grade 7 mathematics teachers in the school had a class selected. Regular and advanced grade 7 mathematics classes were both eligible for observation, and the random selection process did not privilege one over the other.

The study team selected the classes to observe before the start of the classroom observation phase of data collection. The district pacing guide indicated that the unit would be taught in February. Starting in mid-January, the study team asked teachers to indicate the dates when they expected to cover statistics. Almost all teachers indicated that the statistics unit would instead start in March, and many of those and other teachers further delayed the start of the unit until April or May. This occurred equally in STEPSS program schools and practice-as-usual schools. Whenever individual teachers provided dates that they expected to teach statistics lessons, the study team sent an observer on the first available date. About 62 percent of observations in STEPSS program schools (24 of 39) and 69 percent of observations in practice-as-usual schools (24 of 35) were conducted in regular grade 7 mathematics classes, and 38 percent of observations in STEPSS program schools and 31 percent of observations in practice-as-usual schools were conducted in advanced grade 7 mathematics classes. The analyses include a course type indicator (regular or advanced) to control for any potential differences between classes (see appendix B).

Leaders in the BCPS curriculum department notified the schools assigned to the STEPSS program of the change in their instructional materials to the STEPSS materials for the 2018/19 school year. The district leaders then invited the teachers who taught grade 7 mathematics in these schools during the 2017/18 school year to participate in the corresponding professional development.

Measuring student understanding of statistics

The study measured one primary student outcome—student understanding of statistics. The study administered the 23-item, paper-pencil, Beginner/Intermediate, Form 1 of the Levels of Conceptual Understanding in Statistics (LOCUS) test to participating students. The National Science Foundation funded the LOCUS project (Jacobbe et al, 2014; DRL-1118168) to develop valid and reliable assessments of students' understanding of statistics. As a part of the development process, subject matter experts reviewed the LOCUS items and assessment framework and judged them to be consistent with the current curriculum standards in probability and statistics found in most schools in the United States. Twenty of the original 23 items on the test form contributed to the final student scores, with the remaining 3 items removed due to poor discrimination estimates using the current study data (Huggins-Manley & Jacobbe, 2019).

LOCUS scale scores are generated by models based on item response theory and transformed to a scale of 0–200. Reliability estimates for the LOCUS tests on this scale have been .90 or higher in feasibility tests (Jacobbe, 2016).

The current study placed the student ability estimates on an alternative scale of 20 to 80, with a mean of 50 and a standard deviation of 10, after the deletion of the three items from the item set. The marginal reliability of the scale scores from the 20-item LOCUS scale was 0.68. McDonald's ω was .74 (Huggins-Manley & Jacobbe, 2019; Zinbarg et al., 2005). Teachers in both groups administered the LOCUS test to their students in either April or May, depending on the timing of the conclusion of the unit, and then transmitted the test forms for students with approved parental consent and student assent to the study team. Appendix B summarizes the LOCUS test data and provides an average treatment effect to answer research question 1 on the effect on student understanding of statistics.

Scores on the statewide grade 7 mathematics assessment for participating students were not used as an outcome measure even though the study team had access to them. In many of the schools, students took the statewide assessment prior to the implementation of the statistics unit, which precluded the use of the assessment as an outcome measure.

Measuring classroom instruction

The study team observed a single day of classroom statistics instruction for each teacher in STEPSS program schools and practice-as-usual schools and rated the observed cognitive complexity and classroom discourse using the Instructional Quality Assessment (IQA). The IQA measures several components of classroom instruction, organized into two broad categories: cognitive complexity of student tasks and classroom discourse (Boston, 2012; Junker et al., 2005). The full IQA instrument comprises nine individual rubrics (see table A2 for a description of each rubric). Each rubric uses a -point Likert-type scale, with a minimum of 0 and a maximum of 4. Observers score all the rubrics after a live classroom observation. In this study, teachers in STEPSS program schools were asked to use the STEPSS lessons, which consistently score higher on the task potential rubric than the lesson plans in the district-adopted textbook, the most common source of instructional materials in practice-as-usual schools. As such, the measure of task potential was directly affected by the curriculum materials provided to STEPSS program schools. Out of concern about potentially inflating the scores, the study team did not include the task potential score in the final IQA score for each teacher. The total IQA score had a range of 0–32 and was calculated by summing the rubric score, rated on a scale of 0–4, for the remaining eight rubrics used in the study.

Observers participated in four days of training that included opportunities to study the applicable curriculum standards for probability and statistics, the lesson plans in the STEPSS units and the district-adopted curriculum, and the IQA and related publications. Observers also practiced conducting observations by scoring videos of classroom instruction using the IQA and reaching consensus with partners. The study team also reviewed the onsite protocol for conducting observations and the data entry protocol with observers. The study team did not indicate the treatment condition of the schools to the observers when they were dispatched to schools, but almost all of the STEPSS program schools used the STEPSS lessons, and it was easy for the observers to guess that the distinction in curriculum materials was a distinguishing feature.

Table A2. Instructional Quality Assessment mathematics rubrics for lesson observations

Rubric	Description
Task potential	The highest level of thinking required by the main instructional task(s) identified as consuming the most instructional time.
Task implementation	The highest level of engagement of the majority of students during their work on and discussion of the main instructional task(s).
Student discussion following task	Highest level of rigor during the final discussion of the lesson, following students' work on the main instructional task.
Teacher questions	The highest level of cognitive processes elicited by the teacher's questions.
Participation	Percentage of students who make verbal contributions at some time the entire lesson.
Teacher linking	Teachers' efforts to prompt students to connect, extend, analyze, or critique the mathematical work and thinking of others during the whole-group discussion following students' work on the main instructional task.
Student linking	Students' efforts to connect, extend, analyze, or critique the mathematical work and thinking of others during the whole-group discussion following students' work on the main instructional task.
Teacher press	Teachers' efforts to prompt students to justify the accuracy of their computations, explain their thinking, and validate their claims.
Student providing	The extent to which students justify the accuracy of their computations, explain their thinking, and validate their claims.

Source: Boston, 2012.

The study team observed 74 lessons. For 53 of these lessons, two study team members observed the lesson, rated each of the IQA rubrics, and then compared their ratings. For the other 21 lessons a single study team member performed the observation. In both cases (that is, two observers and one observer), a single IQA score for each observation was used. When two observers compared their ratings, they discussed any disagreements, used evidence based on their notes, and agreed on a final rating for each rubric. This approach results in an ill-structured measurement design in which the ratings are neither fully crossed nor nested for the purpose of calculating interrater reliability. Interrater reliability under this design can be estimated using the reliability estimator $G(q,k)$, which is equally appropriate for crossed, nested, and ill-structured designs (Putka et al., 2008). Using $G(q,k)$, the interrater reliability of all independent ratings prior to any consensus coding is 0.804. Appendix B summarizes the data collected through these observations and provides average treatment effects to answer research question 2 on the effect on statistics instruction.

Fidelity of implementation

The STEPSS program comprises two major components: professional development workshops and curriculum materials. This section provides data on fidelity of implementation of each component.

Participation in professional development workshops. Teacher participation in the professional development workshops is an important aspect of implementation fidelity for the STEPSS program. The study team kept attendance records for each workshop. Almost all eligible teachers in the participating schools assigned to the STEPSS program participated in the workshops (table A3). All of the teachers in the analytic sample for the analysis of statistics instruction participated in at least two days of workshops, and 97 percent of the analytic sample participated in each training opportunity. A lower percentage (89 percent) of teachers of students in the analytic sample for the analysis of student understanding of statistics participated in at least two days of workshops. Post hoc analyses of the reasons for nonparticipation indicated that scheduling conflicts (for example, illness, maternity leave, failure of substitute teachers to report for duty) and teacher turnover during the school year explained almost all of the deficit.

Table A3. Number of teachers in the analytic sample for the Supporting Teacher Enactment of the Probability and Statistics Standards program schools who attended professional development workshops, by outcome variable, Broward County Public Schools, 2018/19

Outcome variable	Number of teachers (and schools) included	Percent participating in training 1	Percent participating in training 2	Percent participating in training 1 or 2
Instructional Quality Assessment	39 (14)	97	97	100
Levels of Conceptual Understanding in Statistics	47 (17)	87	81	89

Note: Training 1 and 2 each consisted of two days of professional development workshops for a total of four days of workshops. Teachers are counted as participating if they attended at least one of the two days at each training.
Source: Authors' compilation.

Implementation of the lesson plans in the STEPSS program replacement unit for probability and statistics. For each observed lesson the observer noted the source of the main lesson plan for the day (table A4). These data indicate that teachers in the STEPSS program schools used the STEPSS program replacement unit during statistics instruction, although 2 of the 39 observed lessons came from the district-adopted curriculum. Fewer than half of the lessons observed in practice-as-usual schools came from the district-adopted curriculum. About 60 percent of the observed lessons in practice-as-usual schools came from sources other than the curriculum materials adopted by the school district. These sources included teacher-created lessons, lessons or activities found by teachers on the Internet through sources such as Khan Academy or Pinterest, lessons drawn from textbooks other than the district-adopted series, and practice problems drawn from the set of released problems from the state mathematics assessment.

Table A4. Number and percentage of lessons drawn from the district-adopted curriculum, the Supporting Teacher Enactment of the Probability and Statistics Standards replacement unit, or other sources, by condition, Broward County Public Schools, 2018/19

	STEPSS program		Practice as usual	
Source	Number	Percent	Number	Percent
District-adopted textbook	2	5	14	40
STEPSS replacement unit	37	95	0	0
Other	0	0	21	60

STEPSS is Supporting Teacher Enactment of the Probability and Statistics Standards.
Note: Observations occurred in 39 lessons in 14 STEPSS program schools and in 35 lessons in 12 practice-as-usual schools.
Source: Authors' analysis of data.

For the STEPSS program the study team used the difference between task potential scores and task implementation scores on the IQA as a fidelity metric (table A5), with each rubric scored on a scale of 0–4. This metric quantifies the achieved cognitive complexity of tasks implemented in the classroom relative to the intended cognitive complexity of the instructional materials and provides insight into whether potential differences in opportunities to learn are resulting from the task potential—a matter of instructional materials—or teacher implementation of those instructional materials. All lessons were scored 3 or 4 on the task potential rubric, with 19 scores of 3 and 20 scores of 4. The task implementation scores consisted of 16 scores of 2, 19 scores of 3, and 4 scores of 4. In 14 of the 39 observations, the task implementation score matched the task potential score. In the other 25 observations the task implementation score was below the task potential score, with 18 observations scored one point lower and 7 observations scored two points lower. The mode difference was one point. None of the task implementation scores exceeded the task potential scores.

Table A5. Number of task potential scores and task implementation scores at each level for lessons observed in schools participating in the Supporting Teacher Enactment of the Probability and Statistics program, Broward County Public Schools, 2018/19

	Task Implementation score			
Task potential score	Score of 2	Score of 3	Score of 4	Total
Score of 3	9	10	0	19
Score of 4	7	9	4	20
Total	16	19	4	39

Note: Includes only observed score levels.
Source: Authors' compilation.

References

Boston, M. (2012). Assessing instructional quality in mathematics. *The Elementary School Journal, 113*(1), 76–104.

Burger, E. B., Dixon, J. K., Kanold, T. D., Larson, M. R., Leinwand, S., & Sandoval-Martinez, M. E. (2016). *GoMath! Middle School.* Houghton Mifflin Harcourt.

Fernández, M. L. (2005). Learning through microteaching lesson study in teacher preparation. *Action in Teacher Education, 26*(4), 37–47.

Fernández, M. L. (2010). Investigating how and what prospective teachers learn through microteaching lesson study. *Teaching and Teacher Education, 26*(2), 351–362.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., et al. (2007). *Guidelines for assessment and instruction in statistics education report: A pre-K–12 curriculum framework.* American Statistical Association.

Hopfensperger, P., Jacobbe, T., Lurie, D., & Moreno, J. (2012). *Bridging the gap between Common Core State Standards and teaching statistics*. American Statistical Association.

Huggins-Manley, C., & Jacobbe, T. (2019). *Psychometric analysis of LOCUS as implemented through the Supporting Teachers' Enactment of the Probability and Statistics Standards (STEPSS) Project* (Research Report). University of Florida.

Jacobbe, T. (2016). *Levels of Conceptual Understanding in Statistics (LOCUS): 2016 test summary report*. University of Florida.

Jacobbe, T., Case, C., Whitaker, D., & Foti, S. (2014). Establishing the content validity of the LOCUS assessments through evidence centered design. In K. Makar & R. Gould (Eds.), *Proceedings of the 9th International Conference on Teaching Statistics*. International Statistical Institute. Retrieved October 28, 2020, from https://icots.info/9/proceedings/home.html.

Junker, B., Matsumura, L. C., Crosson, A., Wolf, M. K., Levison, A., Weisberg, Y, et al. (2005, April). *Overview of the Instructional Quality Assessment*. Paper presented at the Annual Meeting of the American Educational Research Association.

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology*, *93*(5), 959.

Zhou, G., & Xu, J. (2017). Microteaching lesson study: An approach to prepare teacher candidates to teach science through inquiry. *International Journal of Education in Mathematics, Science and Technology, 5*(3), 235–247.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123–133.

Appendix B. Methods

The study team used multilevel statistical analysis to estimate the effect of the Supporting Teacher Enactment of the Probability and Statistics Standards (STEPSS) program on student understanding of statistics and on statistics instruction. This appendix provides details on the study data included in these analyses as well as the specific statistical models used to estimate the effects of the STEPSS program.

Study data

The initial sample for research question 1 on the effect on student understanding of statistics included all 40 eligible middle schools in Broward County Public Schools (BCPS). The study team randomly selected 20 schools from this sample to participate in the STEPSS program and assigned the remaining 20 schools to practice as usual. The randomized sample includes all teachers of regular and advanced grade 7 mathematics in these schools at the beginning of the 2018/19 school year and all students listed on the regular and advanced grade 7 mathematics rosters during the school district's census survey of students in October. At baseline the study sample included 6,911 grade 7 mathematics students in the STEPSS program and 7,134 grade 7 mathematics students in practice-as-usual schools for a total of 14,045 eligible students.

The analytic sample sustained a substantial loss of schools, teachers, and students. The approved consent process required active, informed parental consent, and the primary source of attrition at the student level was nonreturn of parental consent forms by the parents (figure B1). Several schools refused to participate in any of the study activities. The final analytic sample includes students who met all of the following criteria: had parental consent and assented to participate in the study, completed the outcome assessment, had baseline data available on the grade 6 Florida Standards Assessment (FSA) Mathematics, and remained in the same school until the Levels of Conceptual Understanding in Statistics (LOCUS) testing window. The overall school-level attrition rate is 23 percent (table B1). Attrition differs by 15 percent between STEPSS schools and practice-as-usual schools. Of the 11,118 eligible students in the 31 schools in the analytic sample, 2,283 students participated in the analysis of the effect on student understanding of statistics, resulting in a student nonresponse rate of 79 percent.

Table B1. Summary of sample attrition from Supporting Teacher Enactment of the Probability and Statistics Standards program schools and practice-as-usual schools for the analysis of effect on student understanding of statistics (research question 1), Broward County Public Schools, 2018/19

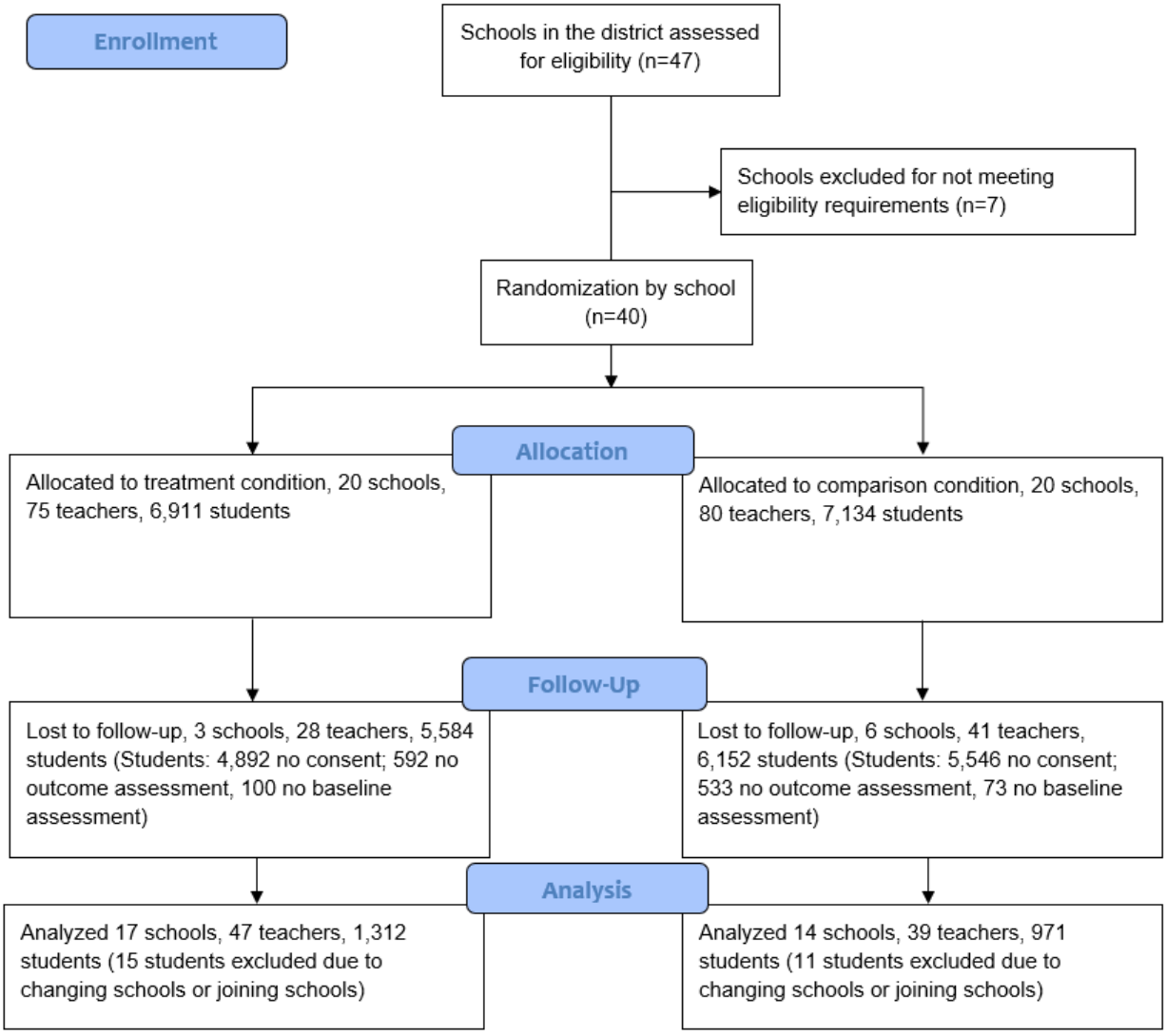
	Number at baseline		Number in the analytic sample		Attrition	
Analysis level	STEPSS program	Practice as usual	STEPSS program	Practice as usual	Overall (percent)	Differential (percent)
Schools	20	20	17	14	23	15

STEPSS is Supporting Teacher Enactment of the Probability and Statistics Standards.

Note: Differential attrition is in absolute values. At baseline 6,911 students were in schools assigned to the STEPSS program and 7,134 students were in schools assigned to practice as usual. In the 17 schools in the STEPSS program analytic sample, 1,312 of the 5,841 eligible students remained in the sample. In the 14 schools in the practice-as-usual analytic sample, 971 of the 5,277 eligible students remained in the sample. The student nonresponse rate was 79 percent.

Source: Authors' compilation.

Figure B1. Sample attrition from Supporting Teacher Enactment of the Probability and Statistics Standards program schools and practice-as-usual schools for the analysis of the effect on student understanding of statistics (research question 1), Broward County Public Schools, 2017/18–2018/19



Source: Authors' compilation using the CONSORT 2010 template.

Following What Works Clearinghouse (WWC) guidelines, the study team assessed baseline equivalence on prior achievement for the analytic group using STEPSS program and practice-as-usual frequencies and unadjusted score means and standard deviations on the grade 6 FSA Mathematics. The effect size difference in achievement between the groups is 0.049 (in absolute value), which is in the acceptable range for demonstrating baseline equivalence in prior achievement (table B2). While no adjustment is necessary to meet WWC standards with reservations, grade 6 achievement scores are included in the estimation of treatment effects as described later in the explanation of the analysis models. Differences in student race/ethnicity and gender are provided for reference only, as race/ethnicity and gender are not included in the analyses.

Table B2. Baseline equivalence for the analytic sample of students in Supporting Teacher Enactment of the Probability and Statistics Standards program schools and practice-as-usual schools, Broward County Public Schools, 2017/18

Characteristic	Students in STEPSS program schools (*n* = 1,312)	Students in practice-as-usual schools (*n* = 971)	Effect size (absolute value)
Average grade 6 score on the Florida Standards Assessment Mathematics	327.24 (21.46)	328.31 (22.21)	0.049
Percentage of Black students	30.5	23.6	0.194
Percentage of Hispanic students	32.0	36.4	0.108
Percentage of White students	30.3	33.8	0.089
Percentage of male students	47.0	44.0	0.067

STEPSS is Supporting Teacher Enactment of the Probability and Statistics Standards.
Note: Values in parentheses are standard deviations. The number of STEPSS program schools is 17, and the number of practice-as-usual schools is 14.
Source: Authors' compilation of data from Broward County Public Schools.

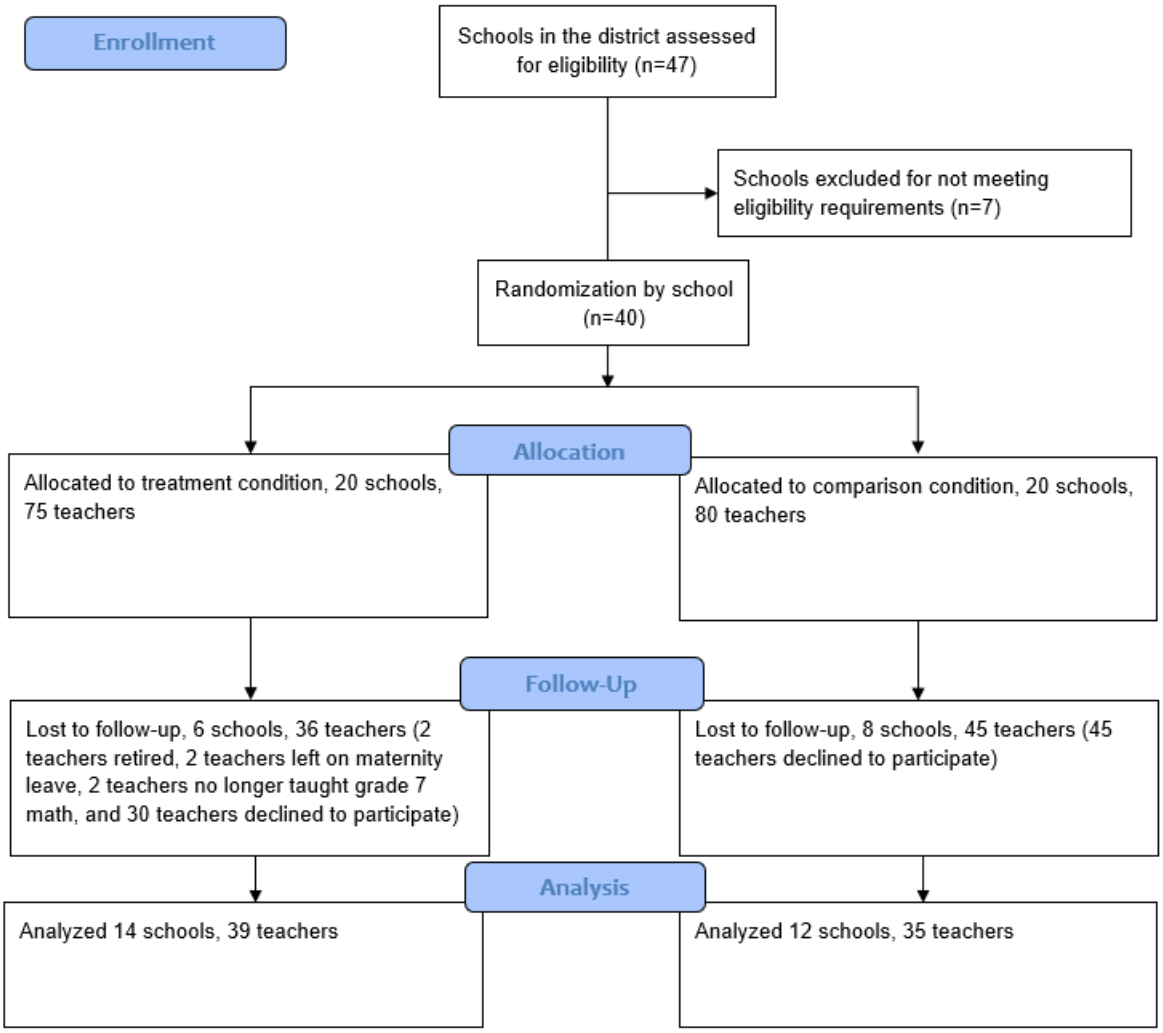
The initial sample for research question 2 on the effect on statistics instruction included all 40 eligible middle schools in BCPS, with the same school-level assignment as in research question 1 and the same number of teachers in those schools. The analytic sample for this outcome included 26 schools, resulting in an overall school-level attrition rate of 35 percent (table B3 and figure B2). Attrition differs by 10 percent between STEPSS program schools and practice-as-usual schools. Attrition occurred primarily due to schools or teachers declining to allow the study team to observe their classrooms. Of the 105 eligible teachers in the 26 schools in the analytic sample, 74 participated in the analysis of the effect on statistics instruction, resulting in a teacher nonresponse rate of 30 percent.

Table B3. Summary of sample attrition from Supporting Teacher Enactment of the Probability and Statistics Standards program schools and practice-as-usual schools for the analysis of the effect on statistics instruction (research question 2), Broward County Public Schools, 2018/19

	Number at baseline		Number in the analytic sample		Attrition	
Analysis level	STEPSS program	Practice as usual	STEPSS program	Practice as usual	Overall (percent)	Differential (percent)
Schools	20	20	14	12	35	10

STEPSS is Supporting Teacher Enactment of the Probability and Statistics Standards.
Note: Differential attrition is in absolute values. At baseline 75 teachers were in schools assigned to the STEPSS program, and 80 teachers were in schools assigned to practice as usual. In the 14 schools in the STEPSS program analytic sample, 39 of the 53 eligible teachers remained in the sample. In the 12 schools in the practice-as-usual analytic sample, 35 of the 52 eligible teachers remained in the sample. The teacher nonresponse rate was 30 percent.
Source: Authors' compilation.

Figure B2. Sample attrition from Supporting Teacher Enactment of the Probability and Statistics Standards program schools and practice-as-usual schools for the analysis of the effects on statistics instruction (research question 2), Broward County Public Schools, 2018/19



Source: Authors' compilation using the CONSORT 2010 template.

Following WWC guidelines, randomized controlled studies with high attrition, as in this study, must demonstrate baseline equivalence to meet WWC standards with reservations. However, this study does not have a baseline measure of statistics instruction or other teacher characteristics to evaluate equivalence or potential bias due to missing data, so the analysis of the effect on statistics instruction is not eligible to meet WWC group design standards and is limited by this expected rating.

The number of teachers in the sample differs for each research question outcome (table B4). These differences may be due in part to the number of teachers declining to participate in teacher observations. However, the sample of teachers with each outcome overlapped substantially, with at least two-thirds having both outcomes.

Table B4. Number of teachers in Supporting Teacher Enactment of the Probability and Statistics Standards program schools and practice-as-usual schools represented in the analytic sample, by outcome variable

Condition	Instructional Quality Assessment	Levels of Conceptual Understanding in Statistics	Instructional Quality Assessment and Levels of Conceptual Understanding in Statistics
STEPSS program	39	47	32
Practice as usual	35	39	26

STEPSS is Supporting Teacher Enactment of the Probability and Statistics Standards.
Note: Values refer to the full sample of teachers with each outcome or outcomes.
Source: Authors' compilation.

Data analysis

The study team estimated the main effect of the STEPSS program on student understanding of statistics (research question 1) using a three-level hierarchical model, with students nested within teachers and teachers nested within schools. The study team conducted the analyses in SAS 9.4 using PROC MIXED. Random assignment to the STEPSS program occurred at the school level, so the treatment effect is included in the level 3 model.

Level 1 (Students): $$LOCUS\ Scale\ Score_{ijk} = \pi_{0jk} + \pi_{1jk} Advanced_{ijk} + \pi_{2jk} Grade\ 6\ FSA_{ijk} + e_{ijk},$$

Level 2 (Teachers): $$\pi_{0jk} = B_{00k} + r_{0jk},$$
$$\pi_{1jk} = B_{10k}$$
$$\pi_{2jk} = B_{20k},$$

Level 3 (Schools): $$B_{00k} = \gamma_{000} + \gamma_{001} STEPSS\ Program_k + \mu_{00k},$$
$$B_{10k} = \gamma_{100}$$
$$B_{20k} = \gamma_{200}$$

where $LOCUS\ Scale\ Score_{ijk}$ is the continuous outcome measure for student *i* of teacher *j* and school *k*, $Advanced_{ijk}$ is an indicator variable taking a value of 1 for students enrolled in advanced grade 7 mathematics and 0 for students enrolled in regular grade 7 mathematics, and $Grade\ 6\ FSA_{ijk}$ is the grand mean–centered scale score on the FSA Mathematics in grade 6 for student *i*. The error term at level 1, e_{ijk} , represents the deviation of student *ijk*'s LOCUS scale score from the predicted score and is assumed to be distributed N(0, σ^2). $STEPSS\ Program_k$ is an indicator variable taking a value of 1 for STEPSS program schools and 0 for practice-as-usual schools, γ_{000} is the covariate-adjusted grand mean, γ_{001} is the treatment effect, and u_{00k} is a level 3 random effect that represents the deviation of school *k*'s intercept from its predicted value. The test for the treatment effect in this design is a test of the hypothesis that γ_{001} = 0.

The treatment effect in the final model (model 4, table B5) is significant (p = .031) and is estimated at 1.85. The least squares means statement in the SAS PROC MIXED procedure was used to calculate adjusted group means. The adjusted mean LOCUS scale score for the treatment group is 50.72, and the adjusted mean LOCUS scale score for the practice-as-usual group is 48.87 (table B6). The mean difference is equal to an effect size of 0.23 based on calculation guidelines recommend by the WWC (U.S. Department of Education, 2017), and an improvement index of +9. The WWC improvement index is the expected change in percentile rank for an average practice-as-usual student if the student had received the treatment.

Table B5. Estimates from the three-level model predicting student understanding of statistics on the Levels of Conceptual Understanding in Statistics test, Broward County Public Schools, 2017/18–2018/19

Model components	Model 1	Model 2	Model 3	Model 4
Fixed effects				
Intercept	48.83*(0.62)	49.85* (0.45)	49.01* (0.51)	48.01* (0.67)
Grade 6 FSA Mathematics score		0.20* (0.01)	0.18* (0.01)	0.18* (0.01)
Advanced grade 7 mathematics course[a]			1.60* (0.43)	1.58* (0.43)
STEPSS program				1.85* (0.85)
Error variance				
Level 1	48.99* (1.48)	37.22* (1.12)	36.99* (1.12)	36.98* (1.11)
Intercept (Teacher)	19.27* (4.41)	11.54* (2.79)	11.75* (2.82)	11.74* (2.82)
Intercept (School)	2.91 (3.14)	0.82 (2.07)	0.71 (2.05)	0.15 (1.92)
Model fit				
Akaike information criterion	15,555.1	14,914.6	14,900.6	14,894.5
Bayesian information criterion	15,549.1	14,918.9	14,904.9	14,898.8

* Statistically significant at $p < .05$.
FSA is Florida Standards Assessment. STEPSS is Supporting Teacher Enactment of the Probability and Statistics Standards.
Note: The school intraclass correlation coefficient is .04, and the teacher intraclass correlation coefficient is .27. Grade 6 FSA Mathematics scores are grand mean centered. Entries show parameter estimates with standard errors in parentheses. The estimation method is restricted maximum likelihood. The total number of students is 2,283 (1,312 students in 17 STEPSS program schools and 971 students in 14 practice-as-usual schools).
a. Advanced is an indicator variable at level 1 taking a value of 1 if the observation occurred in an advanced grade 7 mathematics class and 0 otherwise. The variable was removed in the final model due to its insignificance.
Source: Authors' calculations using data from Broward County Public Schools.

Table B6. Means, unadjusted standard deviations, and effect size estimates for the complete case sample, Broward County Public Schools, 2017/18–2018/19

	Number of schools		Adjusted mean (unadjusted standard deviation)		What Works Clearinghouse calculations		
Measure	STEPSS program	Practice as usual	STEPSS program	Practice as usual	Mean difference	Effect size	Improvement index
LOCUS	17	14	50.72 (8.31)	48.87 (8.07)	1.85	0.23	9

LOCUS is Levels of Conceptual Understanding in Statistics. STEPSS is Supporting Teacher Enactment of the Probability and Statistics Standards.
Note: The analytic sample includes 1,312 students in STEPSS program schools and 971 students in practice-as-usual schools.
Source: Authors' compilation.

The study team estimated the main effects of the STEPSS program on statistics instruction (research question 2) using a two-level hierarchical model, with teachers nested within schools. The study team conducted the analyses in SAS 9.4 using PROC MIXED. Random assignment to the treatment condition occurred at the school level, so the treatment effect is included in the level 2 model.

Level 1 (Teachers): $$IQA\ Total\ Score_{ij} = B_{0j} + r_{ij},$$

Level 2 (Schools): $$B_{0j} = \gamma_{00} + \gamma_{01} STEPSS\ Program_j + \mu_{0j},$$

where $IQA\ Total\ Score_{ij}$ is the continuous outcome measure for teacher *i* of school *j* and is the sum of the final ratings on all Instructional Quality Assessment (IQA) rubrics except task potential. B_{0j} is the mean IQA score of teachers in school *j*, and r_{ij} is the random error associated with teacher *i* in school *j* and is assumed to be distributed N(0, σ^2). $STEPSS\ Program_j$ is an indicator variable taking a value of 1 for STEPSS program schools and 0 for practice-as-usual schools, γ_{00} is the mean IQA score in practice-as-usual schools, γ_{01} is the treatment effect, and μ_{0j} is the random error associated with school *j* and is independently normally distributed with mean 0 and variance τ_{00}. The test for the treatment effect in this design is a test of the hypothesis that γ_{01} = 0.

The treatment effect in the final model (model 3, table B7) is statistically significant (p = .003) and is estimated at 3.77. The adjusted mean IQA total score for the treatment group is 19.38, and the adjusted mean IQA total score

for practice as usual is 15.61 (table B8). The mean difference is equal to an effect size of 0.80 based on calculation guidelines recommend by the WWC (U.S. Department of Education, 2017), and an improvement index of +29. A positive effect of this magnitude indicates that the STEPSS program is likely to change classroom instruction. However, because the analysis of this outcome is on a random sample with high attrition and without pretreatment classroom instruction data available to compare the two groups for baseline equivalence, there is a possibility of existing contrary evidence not captured in the sample data. Due to this possibility, readers should interpret the findings with caution.

Table B7. Estimates from the two-level model predicting classroom instruction as measured by the Instructional Quality Assessment, Broward County Public Schools, 2018/19

Model components	Model 1	Model 2	Model 3
Fixed effects			
Intercept	17.68* (0.71)	17.15* (0.80)	15.61* (0.88)
Advanced grade 7 mathematics course[a]		1.47 (1.16)	
STEPSS program			3.77* (1.21)
Error variance			
Level 1	19.80* (4.02)	20.07* (4.12)	19.87* (4.03)
Intercept (School)	5.60 (3.93)	4.88 (3.82)	2.17* (3.05)
Model fit			
Akaike information criterion	447.8	444.1	437.1
Bayesian information criterion	450.3	446.6	439.6

* Statistically significant at $p < .05$.
STEPSS is Supporting Teacher Enactment of the Probability and Statistics Standards.
Note: School intraclass correlation coefficient is .22. Entries show parameter estimates with standard errors in parentheses. The estimation method is restricted maximum likelihood. The total number of teachers is 74 (39 teachers in 14 STEPSS program schools and 35 teachers in 12 practice-as-usual schools.
a. Advanced is an indicator variable at level 1 taking a value of 1 if the observation occurred in an advanced grade 7 mathematics class and 0 otherwise. The variable was removed in the final model due to its insignificance.
Source: Authors' calculations using data from Broward County Public Schools.

Table B8. Effect size estimate of the impact of the Supporting Teacher Enactment of the Probability and Statistics Standards program on IQA total score, Broward County Public Schools, 2018/19

	Number of schools		Adjusted mean (unadjusted standard deviation)		What Works Clearinghouse calculations		
Measure	STEPSS program	Practice as usual	STEPSS program	Practice as usual	Mean difference	Effect size	Improvement index
IQA	14	12	19.38 (5.10)	15.61 (4.16)	3.77	0.80	29

IQA is Instructional Quality Assessment. STEPSS is Supporting Teacher Enactment of the Probability and Statistics Standards.
Note: The analytic sample includes 39 teachers in STEPSS program schools and 35 teachers in practice-as-usual schools.
Source: Authors' compilation.

Reference

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2017). *What Works Clearinghouse Standards Handbook Version 4.0.*