



# WWC Intervention Report

A summary of findings from a systematic review of the evidence



Adolescent Literacy

February 2018

## Achieve3000®

### Intervention Description<sup>1</sup>

*Achieve3000*® is a supplemental online literacy program that provides nonfiction reading content to students in grades preK–12 and focuses on building phonemic awareness, phonics, fluency, reading comprehension, vocabulary, and writing skills.

*Achieve3000*® is designed to help students advance their nonfiction reading skills by providing differentiated online instruction. Teachers use the program with an entire class, but the assignments are tailored to each student's reading ability level. For example, teachers assign an article and related activities to an entire class; the program then tailors the version of the article to each student by automatically increasing the difficulty of text when a student is ready for more challenging text. Assignments follow a five-step literacy model: (1) respond to a Before Reading Poll, (2) read an article, (3) answer activity questions, (4) respond to an After Reading Poll, and (5) answer a Thought Question. Progress reports and student usage data, provided by the online tool, enable teachers to track both whole-class and individual student progress.

The intervention is designed for diverse student groups, including general education students, struggling readers in need of intensive tutoring, and English learners.

### Research<sup>2</sup>

The What Works Clearinghouse (WWC) identified three studies of *Achieve3000*® that both fall within the scope of the Adolescent Literacy topic area and meet WWC group design standards. Three studies meet WWC group design standards with reservations, and no studies meet WWC group design standards without reservations. Together, these studies included about 32,266 students in grades 2–8 in three school districts in California, New Jersey, and North Carolina<sup>3</sup>.

According to the WWC review, the extent of evidence for *Achieve3000*® on the reading achievement outcomes of adolescent readers was medium to large for two student outcome domains—comprehension and general literacy achievement. No studies meet WWC group design standards in the two other domains, so this intervention report does not report on the effectiveness of *Achieve3000*® for those domains.<sup>4</sup> (See the Effectiveness Summary on p. 6 for more details of effectiveness by domain.)

### Effectiveness

*Achieve3000*® had potentially positive effects on comprehension and general literacy achievement for adolescent readers.

### Report Contents

Overview	p. 1
Intervention Information	p. 3
Research Summary	p. 5
Effectiveness Summary	p. 6
References	p. 8
Research Details for Each Study	p. 9
Outcome Measures for Each Domain	p. 14
Findings Included in the Rating for Each Outcome Domain	p. 15
Supplemental Findings for Each Outcome Domain	p. 17
Endnotes	p. 19
Rating Criteria	p. 20
Glossary of Terms	p. 21

This intervention report presents findings from a systematic review of *Achieve3000*® conducted using the WWC Procedures and Standards Handbook (version 3.0) and the Adolescent Literacy review protocol (version 3.0).

**Table 1. Summary of findings<sup>5</sup>**

Outcome domain	Rating of effectiveness	Improvement index <i>(percentile points)</i>		Number of studies	Number of students	Extent of evidence
		Average	Range			
<b>Comprehension</b>	Potentially positive effects	+6	0 to +11	2	12,698	Medium to large
<b>General literacy achievement</b>	Potentially positive effects	+3	+2 to +3	2	32,110	Medium to large

### Intervention Information

#### Background

The developer and distributor of *Achieve3000*® is Achieve3000, Inc. The current Achieve3000 literacy product line includes *Smarty Ants*® (grades preK–2), *KidBiz3000*® (grades 2–5), *TeenBiz3000*® (grades 6–8), *Empower3000*® (grades 9–12), and *Spark3000*® (adult education). In addition, *KidBizPro*®, *TeenBizPro*®, and *EmpowerPro*® deliver literacy instruction in content-area classrooms (that is, social studies and science). The WWC refers to all of these products as *Achieve3000*® in this intervention report, unless the original study noted the version. Address: 1985 Cedar Bridge Avenue, Suite 3, Lakewood, NJ 08701. Email: [office@achieve3000.com](mailto:office@achieve3000.com). Web: <http://www.achieve3000.com/> Telephone: (888) 968-6822.

#### Intervention details

The *KidBiz3000*®, *TeenBiz3000*®, and *Empower3000*® products, which cover the grade levels applicable to the Adolescent Literacy review, are all delivered through an online instructional format. All students begin by taking a placement test (the LevelSet™ assessment) to identify a student’s reading level. Teachers can assign one non-fiction article to the whole class to read, but the program automatically selects a version of the article aligned to each student’s reading ability based on the placement test results (or prior performance). Instructional materials, which include more than 15,000 nonfiction articles, focus on high-interest, real-world issues. The program monitors individual reading performance and increases the difficulty of the articles as the student’s reading ability improves. Classes follow a five-step instruction routine:

*Step 1: Respond to a Before Reading Poll.* Students start each lesson by taking a poll to express their opinions related to the subject of the article. They answer a multiple-choice question and write an explanation of why they answered the poll’s question as they did.

*Step 2: Read an Article.* Students read a nonfiction article that discusses a contemporary issue and then review a Dig Deeper section that provides them with additional background and details about the content area. A student receives one of 12 English versions or eight Spanish versions of the article matched to his or her reading level. The program defines select vocabulary words, and an audio clip properly pronounces each word.

The program includes a Reading Connections section that provides other tools and tasks to help students comprehend what they read, such as (1) highlighting text for future reference; (2) a note field that enables students to summarize their thoughts on the article or to take notes; (3) a place for students to pose questions about the article’s content; and (4) a field that encourages students to identify key themes in the article. Students are expected to complete, at a minimum, two Reading Connections per lesson.

*Step 3: Answer Activity Questions.* After reading the article, students respond to a series of vocabulary and reading comprehension questions (that is, around summarization, central ideas and details, and text structure and development). Based on responses to these questions, the program determines when students are ready for more complex text and then automatically adjusts the reading level of the text they receive in the next lesson. Teachers can review students’ responses to each question, including whether they chose the correct answer on the first or second try.

*Step 4: Respond to an After Reading Poll.* Students return to the poll question (Step 1) to again express their opinions, factoring in any new information they might have acquired from the article they read. This step aims to teach students the importance of evidence and provides an opportunity to share and reflect on their learning.

*Step 5: Answer a Thought Question.* In the final step of the lesson, students provide a written response to a question based on the article they read in Step 2 that includes examples, reasons, and evidence to support their responses.

In addition, each lesson includes a Stretch Article that is written at a level higher than the student's instructional level. Students can complete this article and accompanying Stretch Activity as homework or immediately after the lesson if he or she finishes early. The Stretch Article can also be embedded into the lesson; students can use the new information they have acquired to revise their answers to the Thought Question.

Professional learning for classroom teachers is available on-site, live online, and via online videos. Sessions include hands-on practice for teachers to master implementation strategies, monitor student data, and create an action plan for each student.

Recommended use is 80 lessons over the course of the school year. *KidBiz3000*<sup>®</sup> has program versions tailored to the standards of each state. The program is accessible on multiple devices and platforms, including Apple, Android, and Chromebook products. Apps enable students to access lessons with or without Internet connectivity from school or from home.

### Cost

Achieve3000, Inc. offers program packages that include software access for the academic school year. As of January 2017, the unit price of \$14,675 covered up to 250 student licenses, 12 teacher licenses, 250 parent/guardian licenses, and 2 days of professional development. The company also offers a per-student pricing option of \$42 with a 100-student minimum. Professional development is required with the per-student option and sold separately at \$2,300 per day. For more information about program options and pricing, contact Achieve3000, Inc. at [office@achieve3000.com](mailto:office@achieve3000.com).

## Research Summary

The WWC identified five eligible studies that investigated the effects of *Achieve3000*® on the reading achievement of adolescent readers. An additional six studies were identified but do not meet WWC eligibility criteria (see the Glossary of Terms in this document for a definition of this term and other commonly used research terms) for review in this topic area. Citations for all 11 studies are in the References section, which begins on p. 8.

The WWC reviewed five eligible studies against group design standards. Three studies are randomized controlled trials or use quasi-experimental designs that meet WWC group design standards with reservations. This report summarizes those three studies. The remaining two studies do not meet WWC group design standards.

**Table 2. Scope of reviewed research**

<b>Grades</b>	2–8
<b>Delivery method</b>	Whole class
<b>Intervention type</b>	Supplement

### Summary of studies meeting WWC group design standards without reservations

No studies of *Achieve3000*® meet WWC group design standards without reservations.

### Summary of studies meeting WWC group design standards with reservations

Borman et al. (2015) conducted a quasi-experimental study that examined the effects of *Achieve3000*® on students in the Chula Vista school district in California. The intervention group consisted of students in grades 4–8 who received *Achieve3000*® in 16 schools. The comparison group consisted of students in grades 4–8 enrolled in demographically similar Chula Vista schools that did not offer *Achieve3000*® and implemented their schools' standard English language arts curriculum. *Achieve3000*® students were matched to comparison students on achievement measures and demographics. The authors demonstrated equivalence of the analytic intervention and comparison group at baseline for the combined sample of students.<sup>6</sup> The WWC based its effectiveness rating on 1-year findings from the combined sample of 9,527 students in grades 4–8: 1,957 students in the intervention group and 7,570 students in the comparison group.

Hill and Lenard (2016) conducted a cluster, or group-based, randomized controlled trial examining the effects of *KidBiz3000*® (the elementary school version of *Achieve3000*®) on students in grades 2–5 in 32 elementary schools in North Carolina. The study matched schools based on end-of-grade reading composite scores from spring 2013. Within each pair of matched schools, the authors randomly assigned one school to receive the intervention and the other to receive business-as-usual literacy instruction. Because the study was a cluster randomized controlled trial that analyzed outcomes for students who enrolled in the school after random assignment, the integrity of the study's random assignment was jeopardized. However, the authors provided evidence of baseline equivalence for the analytic sample.<sup>7</sup> The study took place over two school years: 2013–14 and 2014–15. The WWC based its effectiveness rating on findings from the combined 2-year sample of 22,583 students: 11,802 students were in the *Achieve3000*® group and 10,781 students were in the comparison group.

Tracey and Young (2004) conducted a quasi-experimental study that examined the effects of *KidBiz3000*® (the elementary school version of *Achieve3000*®) on fifth-grade students in five schools in the Bayonne school district in New Jersey. Students were pretested using a standardized assessment administered by the school district. The study compared students receiving *KidBiz3000*® instruction in seven classrooms with students in four classrooms who received business-as-usual literacy instruction.<sup>8</sup> The study took place over one school year (2003–04). The WWC based its effectiveness rating on findings from 156 students in grade 5: 80 students in the four *KidBiz3000*® classrooms with differentiated instruction and 76 students in the four comparison classrooms.

Effectiveness Summary

The WWC review of *Achieve3000*® for the Adolescent Literacy topic area includes student outcomes in four domains: comprehension, general literacy achievement, reading fluency, and alphabetics. The three studies of *Achieve3000*® that meet WWC group design standards reported findings in two of the four domains: comprehension and general literacy achievement. The following findings present the authors’ estimates and WWC-calculated estimates of the size and statistical significance of the effects of *Achieve3000*® on adolescent readers. Additional comparisons are available as supplemental findings in Appendix D. The supplemental findings do not factor into the intervention’s rating of effectiveness. For a more detailed description of the rating of effectiveness and extent of evidence criteria, see the WWC Rating Criteria on p. 20.

Summary of effectiveness for the comprehension domain

Table 1. Rating of effectiveness and extent of evidence for the comprehension domain

Rating of effectiveness	Criteria met
<b>Potentially positive effects</b> <i>Evidence of a positive effect with no overriding contrary evidence.</i>	In the two studies that reported findings, the estimated impacts of <i>Achieve3000</i> ® on outcomes in the <i>comprehension</i> domain were (1) a substantively important positive effect and (2) an indeterminate effect.
Extent of evidence	Criteria met
<b>Medium to large</b>	Two studies that included 12,698 students in 36 schools reported evidence of effectiveness in the <i>comprehension</i> domain.

Two studies that meet WWC group design standards with reservations reported findings in the comprehension domain.

Hill and Lenard (2016) reported, and the WWC confirmed, no statistically significant effects of *Achieve3000*® for students in grades 4 and 5 on the reading composite score of the North Carolina End-of-Grade tests in spring 2014 and 2015. The average effect size across the two school years was not large enough to be substantively important (that is, an effect size of at least 0.25). The WWC characterizes these study findings as an indeterminate effect.

Tracey and Young (2004) reported statistically significant effects of *Achieve3000*® for students in grade 5 on the Scholastic Reading Inventory Assessment. After adjusting for clustering of students within classrooms, the WWC found that this effect was not statistically significant, but was large enough to be considered substantively important according to WWC criteria. The WWC characterizes this study finding as a substantively important positive effect.

Thus, for the comprehension domain, one study reported a substantively important positive effect and one study reported an indeterminate effect. This results in a rating of potentially positive effects, with a medium to large extent of evidence.

Summary of effectiveness for the general literacy achievement domain

Table 2. Rating of effectiveness and extent of evidence for the general literacy achievement domain

Rating of effectiveness	Criteria met
<b>Potentially positive effects</b> <i>Evidence of a positive effect with no overriding contrary evidence.</i>	In the two studies that reported findings, the estimated impacts of <i>Achieve3000</i> ® on outcomes in the <i>general literacy achievement</i> domain were (1) a statistically significant positive effect and (2) an indeterminate effect.
Extent of evidence	Criteria met
<b>Medium to large</b>	Two studies that included 32,110 students in dozens of schools reported evidence of effectiveness in the <i>general literacy achievement</i> domain.

Two studies that met WWC group design standards with reservations reported findings in the general literacy achievement domain.

Borman et al. (2016) reported, and the WWC confirmed, a statistically significant positive effect of *Achieve3000*<sup>®</sup> for students in grades 4–8 on the California Standards Test English Language Arts. The WWC characterizes these study findings as a statistically significant positive effect.

Hill and Lenard (2016) reported, and the WWC confirmed, a statistically significant positive effect of *Achieve3000*<sup>®</sup> for students in grades 2–5 on the LevelSet Lexile score in spring 2015. The authors also reported, and the WWC confirmed, a statistically significant negative effect of *Achieve3000*<sup>®</sup> for students in grades 2–5 on the LevelSet Lexile score in spring 2014. The reported average positive effect across the two school years was neither statistically significant nor large enough to be substantively important according to WWC criteria. The WWC characterizes these study findings as an indeterminate effect.

Thus, for the general literacy achievement domain, one study reported a statistically significant positive effect and one study reported an indeterminate effect. This results in a rating of potentially positive effects, with a medium to large extent of evidence.

### References

#### Studies that meet WWC group design standards without reservations

None.

#### Studies that meet WWC group design standards with reservations

Borman, G. D., Park, S. J., & Min, S. (2015). *The district-wide effectiveness of the Achieve3000 program: A quasi-experimental study*. Madison: Measured Decisions, Inc. and the University of Wisconsin-Madison. Retrieved from <https://eric.ed.gov/?id=ED558845>

Hill, D. V., & Lenard, M. A. (2016). *The impact of Achieve3000 on elementary literacy outcomes: Randomized control trial evidence, 2013-14 to 2014-15*. (DRA Report No. 16.02). Cary, NC: Wake County Public School System, Data and Accountability Department.

**Additional source:**

Hill, D. V., Lenard, M. A., & Page, L. C. (2016, March). *The impact of Achieve3000 on elementary literacy outcomes: Evidence from a two-year randomized control trial*. Paper presented at the Society for Research on Educational Effectiveness (SREE) Spring Conference, Washington, DC. Retrieved from <https://eric.ed.gov/?id=ED567483>

Tracey, D. H., & Young, J. W. (2004). *Evaluation of KidBiz3000: Bayonne study final report*. Lakewood, NJ: Achieve3000

**Additional source:**

Tracey, D. H., & Young, J. W. (2005). *Bayonne, NJ schools 2003-2004*. Lakewood, NJ: Achieve3000.

#### Studies that do not meet WWC group design standards

Shannon, L., & Grant, B.-J. (2015). *A final report for the evaluation of the Achieve3000 programs*. Charlottesville, VA: Magnolia Consulting. Retrieved from <https://eric.ed.gov/?id=ED563449> The study does not meet WWC group design standards because the analysis does not provide a credible measure of the effectiveness of the intervention.

**Additional source:**

Magnolia Consulting. (2017). *An addendum to the final report for the evaluation of the Achieve3000 programs: Impact analyses for grades 6 and 9*. Charlottesville, VA: Author.

Young, J. W. (2005). *Evaluation study: Chicago Public Schools 2003-2004*. Lakewood, NJ: Achieve3000. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

#### Studies that are ineligible for review using the Adolescent Literacy Evidence Review Protocol

Achieve3000. (2009). *National Study: Lexile® analysis reveals dramatic non-fiction reading skill increases, 2009*. Lakewood, NJ: Author. The study is ineligible for review because it does not use an eligible design.

Achieve3000. (2011). *National Lexile study: 2009-2010*. Lakewood, NJ: Author. The study is ineligible for review because it does not use an eligible design.

Achieve3000. (2016). *National Lexile Study: 2015-2016*. Lakewood, NJ: Author. The study is ineligible for review because it does not use an eligible design.

Banister-Doughty, S., Hollingsworth, G., & Tittle, F. (2015). *Effective literacy interventions at two middle Tennessee high schools* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3733242) The study is ineligible for review because it does not use an eligible design.

Urdegar, S. M. (2014). *Achieve 3000: An analysis of usage and impact, 2013-2014. Technical Note. Volume 3, Number 1*. Miami, FL: Miami-Dade County Public Schools, Office of Assessment, Research, and Data Analytics. Retrieved from <https://eric.ed.gov/?id=ED561897> The study is ineligible for review because it does not use a sample aligned with the protocol.

Wright, R. (2010). *Impact of a web-based reading program on sixth-grade English language learners* (Doctoral dissertation). Retrieved from <https://eric.ed.gov/?id=ED546286> The study is ineligible for review because it does not use an eligible design.

**Appendix A.1: Research details for Borman et al. (2015).**

Borman, G. D., Park, S. J., & Min, S. (2015). *The district-wide effectiveness of the Achieve3000 program: A quasi-experimental study*. Retrieved from <https://eric.ed.gov/?id=ED558845>

**Table A1. Summary of findings**

**Meets WWC group design standards without reservations**

Outcome domain	Sample size	Study findings	
		Average improvement index (percentile points)	Statistically significant
General literacy achievement	9,527 students	+2	Yes

**Setting** The study was conducted in the Chula Vista school district in California.

**Study sample** This study examined the effects of *Achieve3000*<sup>®</sup> in 16 schools in the Chula Vista school district and used demographically similar schools in the same school district as a comparison group. The study took place over one school year (2011–12) and included students in grades 4–8. The authors matched *Achieve3000*<sup>®</sup> students and comparison students on achievement measures and demographics.

The study participants were 69% Hispanic, 14% Asian, 12% White, and 3% African American. About 30% of students were English learners. About 49% of the study participants received free or reduced-price lunch, and females made up 51% of the sample. The analytic sample included 9,527 students: 1,957 students were in the *Achieve3000*<sup>®</sup> group, and 7,570 students were in the comparison group.

**Intervention group** Intervention students received the *Achieve3000*<sup>®</sup> program over one school year. Most intervention students completed, on average, one or two *Achieve3000*<sup>®</sup> activities per week. The *Achieve3000*<sup>®</sup> intervention was used to supplement a standard core English language arts curriculum. The authors did not identify which core curriculum was used or provide any other information about the implementation of the *Achieve3000*<sup>®</sup> intervention.

**Comparison group** The comparison students did not have access to the *Achieve3000*<sup>®</sup> program. They received English language arts instruction as usual.

**Outcomes and measurement** Outcomes were measured in spring of 2012, and the pretest was administered in the spring of 2011. The authors used students’ English Language Arts (ELA) scores on the California Standards Test (CST) for the combined sample of students in grades 4–8. The outcome was reviewed in the general literacy achievement domain. For a more detailed description of these outcome measures, see Appendix B.

The study presented findings separately by grade. These supplemental findings are reported in Appendix D and do not factor into the intervention’s rating of effectiveness.

**Support for implementation** Schools implementing *Achieve3000*<sup>®</sup> purchased Professional Learning Services along with the programs. These services included 1–3 days per year of on-site support by trainers, including training for teachers new to the school, familiarizing returning teachers with upgrades to the program, one-on-one consulting with teachers, and modeling of best practices.

**Appendix A.2: Research details for Hill and Lenard (2016)**

Hill, D. V., & Lenard, M. A. (April 2016). *The impact of Achieve3000 on elementary literacy outcomes: Randomized control trial evidence, 2013-14 to 2014-15* (DRA Report No. 16.02). Wake County Public School System, Data and Accountability Department.

*Additional source:*

Hill, D. V., Lenard, M. A., & Page, L. C. (2016, March). *The impact of Achieve3000 on elementary literacy outcomes: Evidence from a two-year randomized control trial*. Paper presented at the Society for Research on Educational Effectiveness (SREE) Spring Conference, Washington, DC. Retrieved from <https://eric.ed.gov/?id=ED567483>

**Table A2. Summary of findings** **Meets WWC Group Design Standards With Reservations**

Outcome domain	Sample size	Study findings	
		Average improvement index (percentile points)	Statistically significant
Comprehension	32 schools/12,542 students	0	No
General literacy achievement	24 schools/22,583 students	+3	No

**Setting** The study was conducted in the Wake County Public School System (WCPSS) in Raleigh, North Carolina. As a countywide district, WCPSS has schools representing suburban, urban, and rural areas.

**Study sample** The authors used a cluster randomized controlled trial design to study the effects of *KidBiz3000*<sup>®</sup> on reading achievement for students in grades 2–5. The study took place over two school years (from 2013–14 to 2014–15) in 32 elementary schools. In summer of 2013, the authors matched pairs of schools on the basis of their average 2013 end-of-grade (EOG) reading composite scores, and then from within each matched pair, randomly assigned one school to the intervention group and one school to the comparison group. In both study years, the same 16 *KidBiz3000*<sup>®</sup> schools and 16 comparison schools participated in the study.

The WWC considers random assignment jeopardized because the analytic sample included students who enrolled in study schools after random assignment. For the general literacy achievement domain, the 2-year combined analysis sample included 22,583 students in grades 2–5 in 24 schools:<sup>9</sup> 11,802 students were in the *Achieve3000*<sup>®</sup> group, and 10,781 students were in the comparison group. For the comprehension domain, the 2-year combined analysis sample included 12,542 students in grades 4–5 in 32 schools: 6,585 students were in the *Achieve3000*<sup>®</sup> group, and 5,957 students were in the comparison group. Because some of the same students were analyzed in both years of the study, these reported sample sizes count some individual students more than once.

No demographic data were available on the analytic study sample; however, in the 32 participating schools, the student population was 51% White, 26% African American, and 19% Hispanic. Moreover, 12% of students had disabilities, 9% of students in study schools were English learners, and 7% of students were academically and intellectually gifted (AIG). Approximately one-third of the district’s students were certified for free or reduced-price lunch.

### Intervention group

*KidBiz3000*<sup>®</sup> was implemented in 30-minute sessions, two times per week over a school year. On initial use, students took a 30-minute test to measure their baseline reading achievement. For each lesson, students followed *Achieve3000*<sup>®</sup>'s five-step literacy routine. Across both years of the study, about 8% of intervention students completed at least 80 lessons (i.e., the developer's recommended dosage), about 21% used 40–79 lessons, about 50% used 1–39 lessons, and 22% of students completed no activities. The *Achieve3000*<sup>®</sup> intervention was used to supplement a standard core reading curriculum; however, the authors did not identify which core curriculum was used.

### Comparison group

The comparison condition was business-as-usual reading instruction. Classrooms in comparison schools did not receive a supplemental curriculum.

### Outcomes and measurement

Outcomes were measured in spring 2014 and 2015, and the pretests were administered at the beginning of the school year, in the fall of 2013 and 2014, respectively. All findings reported in the study reflect the impact of the intervention after 1 year of student exposure; in particular, while some students whose outcomes were analyzed in the second year of the study had received the intervention in both years, outcomes from spring 2015 were analyzed using a fall 2014 pretest (which was administered a year after the intervention had begun). The authors used the *Achieve3000*<sup>®</sup> *LevelSet Lexile* assessment for students in grades 2–5. This outcome was reviewed in the general reading achievement domain. The North Carolina EOG test for students in grades 4–5 was reviewed in the comprehension domain. For a more detailed description of these outcome measures, see Appendix B.

The authors also administered the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Oral Reading Fluency assessment for students in grades 2–3. These analyses were reviewed under the Beginning Reading topic area.

The authors present treatment on the treated (TOT) estimates of *Achieve3000*<sup>®</sup> impacts on the study outcomes for each sample. These findings do not meet WWC complier average causal effect (CACE) guidance since the study is a cluster RCT that includes joiners (i.e., students who joined the sample after randomization took place).

The study also presented supplemental findings for a subgroup of academically and intellectually gifted (AIG) students. These supplemental findings are reported in Appendix D and do not factor into the intervention's rating of effectiveness.

The authors also conducted subgroup analyses by special education status (students with disabilities) and English learners. These subgroup analyses are not eligible for review under the Adolescent Literacy review protocol.

### Support for implementation

The study included professional development to train teachers, consisting of two 2.5-hour large group training sessions, and one 1-hour small group session. Teachers were able to obtain follow-up help if needed.

**Appendix A.3: Research details for Tracey & Young (2004).**

Tracey, D. H., & Young, J. W. (2004). *Evaluation of KidBiz3000: Bayonne study final report*. Lakewood, NJ: Achieve3000.

*Additional source:*

Tracey, D. H., & Young, J. W. (2005). *Bayonne, NJ schools 2003-2004*. Lakewood, NJ: Achieve3000.

**Table A3. Summary of findings**

**Meets WWC Group Design Standards With Reservations**

Outcome domain	Sample size	Study findings	
		Average improvement index (percentile points)	Statistically significant
Comprehension	8 classrooms/156 students	+11	No

**Setting** The study was conducted in the Bayonne school district in New Jersey.

**Study sample** This study examined the effects of *KidBiz3000*<sup>®</sup> in 11 fifth-grade classrooms in five Bayonne schools. The study took place over one school year (2003–04). *KidBiz3000*<sup>®</sup> was implemented in seven classrooms: four received differentiated instruction, and three received undifferentiated instruction (see below). Four classrooms participated in the comparison condition.

Students were pretested using a standardized assessment administered by the school district; however, no demographic data were available on the study sample. In the Bayonne school district, the population was 70% White, 18% Hispanic, and 6% African American. One-third of the district’s students were certified for free or reduced-price lunch.

The study established baseline equivalence between the differentiated instruction and comparison classrooms. The analytic sample included 156 students in grade 5: 80 students were in the *KidBiz3000*<sup>®</sup> group, and 76 students were in the comparison group. Baseline equivalence was not established between the undifferentiated instruction and comparison groups, and therefore not included in this review.

**Intervention group** *KidBiz3000*<sup>®</sup> is the elementary school version of *Achieve3000*<sup>®</sup> that targets students in grades 2–5. *KidBiz3000*<sup>®</sup> was implemented in 40-minute sessions in the computer lab, two times per week over the school year. Some students had access to the program from home as well.

The study included two variants of the *KidBiz3000*<sup>®</sup> program, which the study authors referred to as differentiated instruction (four classrooms) and undifferentiated instruction (three classrooms). The differentiated (general) version of the intervention, which uses the program’s full set of features, used materials targeted towards the reading level of the individual student using the intervention software, while the undifferentiated (and unique to this study) version included materials targeted toward the general reading level of the grade.<sup>10</sup>

The *KidBiz3000*<sup>®</sup> intervention was used to supplement a standard core reading curriculum; however, the authors did not identify which core curriculum was used.

### Comparison group

The comparison condition was business-as-usual reading instruction. As standard in the district, they had one session in the computer lab per week. The comparison classrooms did not receive a supplemental curriculum.

### Outcomes and measurement

Outcomes were measured in June 2004, and the pretest was administered in September 2003. The authors used the Scholastic Reading Inventory (SRI) assessment for students in grade 5. The outcome was reviewed in the comprehension domain. For a more detailed description of this outcome measure, see Appendix B.

Supplemental interim SRI findings are presented for students after 3 months (Time 2) and 6 months (Time 3) of exposure to the *KidBiz3000*® intervention. These supplemental findings are reported in Appendix D and do not factor into the intervention's rating of effectiveness.

Reading Composite and Language Composite subtests of the TerraNova (2nd edition, 2000) were also administered in the study. Analyses based on the TerraNova did not meet WWC group design standards because the study did not establish baseline equivalence for the intervention and comparison groups.

The authors also administered the Elementary Reading Attitude Survey, a student survey on attitudes about reading. This outcome is not eligible for review under the Adolescent Literacy review protocol (version 3.0).

### Support for implementation

The study included professional development for teachers, consisting of two 2.5-hour large-group training sessions and one 1-hour small-group session. Teachers were able to obtain follow-up help if needed.

### Appendix B: Outcome measures for each domain

Comprehension	
<i>North Carolina End-of-Grade (EOG) Lexile, Reading Composite score</i>	This standardized state assessment is aligned to the North Carolina English Language Arts Standard Course of Study (Content Standards). Reading comprehension is assessed by having students read authentic selections and then answer multiple-choice questions directly related to the selections. Knowledge of vocabulary is assessed indirectly through application and understanding of terms within the context of selections and questions. The authentic selections in the reading tests are chosen to reflect reading for various purposes such as literary experience, gaining information, and performing a task. The North Carolina Department of Public Instruction contracted with MetaMetrics, Inc. to convert EOG scores into lexiles. E. Lexile levels range from 150 to 1,350 points (as cited in Hill & Lenard, 2016).
Reading comprehension	
<i>Scholastic Reading Inventory (SRI)</i>	The SRI is a diagnostic computer-adaptive test that measures reading comprehension in Lexile units. The test asks students to read passages and answer multiple choice questions to test their understanding of the text. The questions gauge students' comprehension of authentic fiction and nonfiction passages, asking them to paraphrase information, draw conclusions, make inferences, identify supporting details, or make generalizations based on information in the passage. According to Scholastic, Inc., the SRI has been normed in a study with a sample larger than 500,000 students (as cited in Tracey & Young, 2004).
General literacy achievement	
<i>California Standards Test, English Language Arts (CST-ELA)</i>	The CST-ELA measures multiple aspects of ELA achievement: word analysis; fluency; systematic vocabulary development; reading comprehension of informational materials; literacy response and analysis; oral and written fluency and conventions; and writing. The Reading Comprehension subtest includes generating and responding to questions, making predictions, and comparing information from several sources. For grades 4–11, the test consists of 75 multiple-choice questions (as cited in Borman et al., 2016).
<i>Achieve3000® LevelSet Lexile</i>	LevelSet is a criterion-referenced computer-adaptive test that measures mastery of reading skills and comprehension of <i>Achieve3000®</i> program materials. <i>Achieve3000®</i> provides five versions of LevelSet. To ensure students complete test items that are appropriate for their reading levels, the program adapts to an easier version of the test if several successive items are missed. Most students receive 30 multiple-choice questions, but some students may receive fewer (or more) depending on how they answer the questions. For the posttest, all students receive the test according to their current Lexile level (rather than grade level). Lexile levels range from 150 to 1,350 points. The test developer, MetaMetrics, reports the internal-consistency reliability (alpha coefficients) from 0.81 to 0.90 for three test forms across four study grades: 2 through 5 (as cited in Hill & Lenard, 2016 and obtained through the author query).

Appendix C.1: Findings included in the rating for the comprehension domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			p-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
<b>Hill and Lenard (2016)<sup>a</sup></b>								
<i>North Carolina End-of-Grade (EOG) Lexile, reading composite score</i>	2014 sample, Grades 4–5	32 schools/ 6,235 students	989.43 (216.25)	991.06 (215.58)	-1.62	-0.01	0	> .05
<i>North Carolina EOG Lexile, reading composite score</i>	2015 sample, Grades 4–5	32 schools/ 6,307 students	991.77 (220.39)	985.11 (222.87)	6.66	0.03	+1	> .05
<b>Domain average for comprehension (Hill &amp; Lenard, 2016)</b>						<b>0.01</b>	<b>0</b>	<b>Not statistically significant</b>
<b>Tracey and Young (2004)<sup>b</sup></b>								
<i>Scholastic Reading Inventory (SRI)</i>	Grade 5	8 classrooms/ 156 students	866.06 (174.27)	817.57 (156.19)	48.49	0.29	+11	< .01
<b>Domain average for comprehension (Tracey &amp; Young, 2004)</b>						<b>0.29</b>	<b>+11</b>	<b>Not statistically significant</b>
<b>Domain average for comprehension across all studies</b>						<b>0.15</b>	<b>+6</b>	<b>na</b>

**Table Notes:** For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual’s percentile rank that can be expected if the individual is given the intervention. The statistical significance of the study’s domain average was determined by the WWC. Some statistics may not sum as expected due to rounding. na = not applicable.

<sup>a</sup> For Hill and Lenard (2016), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The p-values presented here were reported in the original study. The adjusted group means, unadjusted standard deviations, and sample sizes were obtained through an author query. Findings from 2014 and 2015 are presented separately since these samples partially overlap (i.e., fourth-grade students in the 2014 sample appear as fifth-grade students in the 2015 sample), and because the 2015 sample used a different point of baseline measurement. Findings from both years (2014 and 2015) reflect 1-year impacts for students. This study is characterized as having an indeterminate effect because the mean effect for the measures in this domain was neither statistically significant nor large enough to be substantively important. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

<sup>b</sup> For Tracey and Young (2004), a correction for clustering was needed and resulted in a WWC-computed p-value of .41 for the SRI outcome; therefore, the WWC does not find the result to be statistically significant. The p-value presented here was reported in the original study for the gain scores analysis (p. 9). The WWC calculated the intervention group mean using a difference-in-differences approach by adding the impact of the intervention (i.e., difference in mean gains between the intervention and comparison groups) to the unadjusted comparison group posttest means. This study is characterized as having a substantively important positive effect because the effect for the measure in this domain is positive and not statistically significant but was large enough to be substantively important. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

Appendix C.2: Findings included in the rating for the general literacy achievement domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			p-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
<b>Borman et al. (2015)<sup>a</sup></b>								
<i>California Standards Test English Language Arts (CST-ELA)</i>	Grades 4–8	9,527 students	379.71 (nr)	337.36 (nr)	2.35	0.04	+2	< .01
<b>Domain average for general literacy achievement (Borman et al., 2015)</b>						<b>0.04</b>	<b>+2</b>	<b>Statistically significant</b>
<b>Hill and Lenard (2016)<sup>b</sup></b>								
<i>Achieve3000<sup>®</sup> LevelSet Lexile</i>	2014 sample, Grades 2–5	24 schools/ 9,732 students	550.13 (326.46)	562.00 (345.11)	–11.86	–0.04	–1	< .05
<i>Achieve3000<sup>®</sup> LevelSet Lexile</i>	2015 sample, Grades 2–5	24 schools/ 12,851 students	577.94 (273.70)	520.72 (280.22)	57.22	0.21	+8	< .01
<b>Domain average for general literacy achievement (Hill &amp; Lenard, 2016)</b>						<b>0.09</b>	<b>+3</b>	<b>Not statistically significant</b>
<b>Domain average for general literacy achievement across all studies</b>						<b>0.06</b>	<b>+3</b>	<b>na</b>

**Table Notes:** For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual’s percentile rank that can be expected if the individual is given the intervention. The statistical significance of the study’s domain average was determined by the WWC. Some statistics may not sum as expected due to rounding. na = not applicable; nr = not reported.

<sup>a</sup> For Borman et al. (2015), the WWC did not need to make corrections for multiple comparisons or to adjust for baseline differences. The WWC did not need to make corrections for clustering, as students were matched using the propensity score procedure, so the unit of assignment was at the same level (student) as the unit of analysis. The p-value and effect size presented here were reported in the original study. The study-reported effect size was computed by dividing the Weighted Least Squares (WLS) regression coefficient by the standard deviation of the outcome (p. 14). The effect size reported here is based on the WLS regression using weights based on students’ probabilities of receiving the intervention considering their measured characteristics and covariates for baseline score, student demographic characteristics (ethnicity, race, gender, socioeconomic status, English learner status), and grade level. The reported intervention group mean is calculated as the comparison group mean (intercept) plus the intervention coefficient after adjusting for other covariates (table 6, p. 15). This study is characterized as having a statistically significant positive effect because the estimated effect is positive and statistically significant. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

<sup>b</sup> For Hill and Lenard (2016), a correction for multiple comparisons was needed but did not affect whether the contrasts were found to be statistically significant. The WWC did not need to make corrections for clustering or to adjust for baseline differences. The p-values presented here were reported in the original study. The adjusted group means, unadjusted standard deviations, and sample sizes were obtained through an author query. Findings from 2014 and 2015 are presented separately since these samples partially overlap (i.e., fourth-grade students in the 2014 sample appear as fifth-grade students in the 2015 sample), and because the 2015 sample used a different point of baseline measurement. Findings from both years (2014 and 2015) reflect 1-year impacts for students. This study is characterized as having an indeterminate effect because the reported mean effect for the measures in this domain was neither statistically significant nor large enough to be substantively important. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

Appendix D.1: Description of supplemental findings for the comprehension domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			p-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
<b>Hill and Lenard (2016)<sup>a</sup></b>								
<i>North Carolina EOG Lexile, reading composite score</i>	AIG 2015 sample, Grades 4–5	32 schools/ 788 students	1,221.25 (117.55)	1,244.40 (nr)	-23.15	nr	nr	.06
<b>Tracey and Young (2004)<sup>b</sup></b>								
<i>SRI, time 2: 3 months</i>	Grade 5	8 classrooms/ 156 students	829.40 (159.48)	820.92 (140.40)	8.48	0.06	+2	.88
<i>SRI, time 3: 6 months</i>	Grade 5	8 classrooms/ 156 students	874.93 (141.50)	871.51 (120.26)	3.42	0.03	+1	.87

**Table Notes:** The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual’s percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding. nr = not reported. AIG = academically and intellectually gifted.

<sup>a</sup> For Hill and Lenard (2016), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The p-value, the adjusted group means, unadjusted standard deviations, and sample sizes presented here were obtained through an author query. For the North Carolina EOG Lexile outcome, the authors reported p-value using the results from a regression model that adjusted for pretest scores, but did not report the information needed to calculate a WWC effect size.

<sup>b</sup> For Tracey and Young (2004), the p-values presented here were calculated by the WWC. A correction for clustering was needed but did not affect whether the contrast was found to be statistically significant. The WWC calculated the intervention group mean using a difference-in-differences approach by adding the impact of the intervention (i.e., difference in mean gains between the intervention and comparison groups) to the unadjusted comparison group posttest means.

Appendix D.2: Description of supplemental findings for the general literacy achievement domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			p-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
<b>Borman et al. (2015)<sup>a</sup></b>								
<i>California Standards Test English Language Arts (CST-ELA)</i>	Grade 4	3,006 students	369.45 (nr)	363.49 (nr)	5.96	0.10	+4	< .01
<i>CST-ELA</i>	Grade 5	3,026 students	389.50 (nr)	388.82 (nr)	0.68	0.01	0	> .05
<i>CST-ELA</i>	Grade 6	3,054 students	378.19 (nr)	379.21 (nr)	-1.02	-0.02	-1	> .05
<i>CST-ELA</i>	Grade 7	258 students	402.83 (nr)	384.88 (nr)	17.95	0.40	+16	< .01
<i>CST-ELA</i>	Grade 8	211 students	382.49 (nr)	374.35 (nr)	8.14	0.17	+7	< .05
<b>Hill and Lenard (2016)<sup>b</sup></b>								
<i>Achieve3000<sup>®</sup> LevelSet Lexile</i>	AIG 2014 sample, Grades 4–5	24 schools/ 889 students	985.28 (197.03)	999.20 (nr)	-13.92	nr	nr	.46
<i>Achieve3000<sup>®</sup> LevelSet Lexile</i>	AIG 2015 sample, Grades 4–5	24 classrooms/ 896 students	983.61 (166.89)	952.40 (nr)	31.21	nr	nr	.01

**Table Notes:** The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual’s percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding. nr = not reported. AIG = academically and intellectually gifted.

<sup>a</sup> For Borman et al. (2015), a correction for multiple comparisons was needed and resulted in a WWC-computed critical p-value of .03 for the CST-ELA measure in grade 8, which is within the authors reported p-value range of < .05; therefore, the WWC does not make a determination about the statistical significance of the effect. The WWC confirms the statistical significance of CST-ELA findings for grade 4 and grade 7. The WWC did not need to make corrections for clustering as students were matched using the propensity score procedure, so the unit of assignment is at the same level (student) as the unit of analysis. The p-values and effect sizes presented here were reported in the original study. Sample sizes of students are approximate and based on the study table 3 (p. 89). The study-reported effect size was computed by dividing the Weighted Least Squares (WLS) regression coefficient by the standard deviation of the outcome (p. 14). The effect size reported here is based on the WLS regression using weights based on students’ probabilities of receiving the intervention considering their measured characteristics and covariates for baseline score, student demographic characteristics (ethnicity, race, gender, socioeconomic status, English learner status), and grade level. The reported intervention group mean is calculated as the comparison group mean (intercept) plus the intervention coefficient after adjusting for other covariates (table 6, p. 15).

<sup>b</sup> For Hill and Lenard (2016), a correction for multiple comparisons was needed but did not affect whether the contrasts were found to be statistically significant. The WWC did not need to make corrections for clustering or to adjust for baseline differences. The p-values, the adjusted group means, unadjusted standard deviations, and sample sizes were obtained through an author query. For the *Achieve3000<sup>®</sup> LevelSet Lexile* outcomes, the author reported p-values using the results from a regression model that adjusted for pretest scores but did not report the information needed to calculate a WWC effect size.

### Endnotes

<sup>1</sup> The descriptive information for this intervention comes from publicly available sources: the program's website <http://www.achieve3000.com> (accessed April 3, 2017), Achieve3000 Lessons and Resources, 2010 Price Sheet, Shannon and Grant (2015), and the EdSurge product review (<https://www.edsurge.com/product-reviews/achieve3000>). The What Works Clearinghouse (WWC) requests distributors review the intervention description sections for accuracy from their perspective. The WWC provided the distributor with the intervention description in April 2017, and the WWC incorporated feedback from the distributor. Further verification of the accuracy of the descriptive information for this intervention is beyond the scope of this review. The WWC published a separate intervention report under the Beginning Reading topic area, which covers earlier grades K–3: <https://ies.ed.gov/ncee/wwc/InterventionReport/692>.

<sup>2</sup> The literature search reflects documents publicly available by February 2017. Reviews of the studies in this report used the standards from the WWC Procedures and Standards Handbook (version 3.0) and the Adolescent Literacy review protocol (version 3.0). The evidence presented in this report is based on available research. Findings and conclusions may change as new research becomes available.

<sup>3</sup> Per the Adolescent Literacy topic area protocol, the current intervention report includes students in grades 2–8. Although adolescent readers are defined as students in grades 4–12, the Adolescent Literacy protocol considers findings on students in lower grades eligible for review when the authors aggregate these findings with those in grades 5 or above and do not present findings separately by grade level. Hill and Lenard (2016) reported findings for a combined sample of students in grades 2–5 for the *Achieve3000® LevelSet Lexile* outcomes; therefore, these findings were reviewed and included in this report. Furthermore, the reported sample size of 32,266 students overestimates the number of unique students. Hill and Lenard (2016) included students who attended study schools in two consecutive school years (2013–14 and 2014–15). For example, some fourth-grade students in the 2013–14 sample appear as fifth-grade students in the 2014–15 sample, and are therefore counted twice in the overall sample size presented in this report.

<sup>4</sup> Please see the Adolescent Literacy review protocol (version 3.0) for a list of all the outcome domains.

<sup>5</sup> For criteria used to determine the rating of effectiveness and extent of evidence, see the WWC Rating Criteria on p. 20. These improvement index numbers show the average and range of individual-level improvement indices for all findings across the studies.

<sup>6</sup> For Borman et al. (2015), the reported baseline data are weighted by the Inverse Probability-of-Treatment Weighting weights (Appendix 6 from the study) because the main analysis uses these weights. According to the WWC guidelines for propensity score matching analyses, there should be consistency in the analytic approaches used to demonstrate equivalence and estimate impacts (Reviewer Guidance for Use with the Procedures and Standards Handbook, version 3.0, updated December 2016, p. 19).

<sup>7</sup> The WWC Reviewer Guidance, for use with the WWC Procedures and Standards Handbook (version 3.0), indicates that if the authors of a cluster randomized controlled trial characterize the intervention as having effects on student scores (rather than only on cluster-level scores), and some students enter clusters after random assignment, then the study must demonstrate equivalence of the analytic intervention and comparison groups at baseline.

<sup>8</sup> Tracey and Young (2004) examined the use of *KidBiz3000®* in seven classrooms: four classrooms received differential instruction (full program implementation) and three received undifferentiated instruction (using materials targeted toward the general reading level of the grade, rather than individual students). Students in four comparison classrooms received business-as-usual literacy instruction. Analyses based on the undifferentiated instruction did not meet WWC group design standards.

<sup>9</sup> For the *Achieve3000® LevelSet Lexile* assessment, 10 *KidBiz3000®* schools and 14 comparison schools constituted the school analytic sample (Hill & Lenard, 2016).

<sup>10</sup> Analyses based on the undifferentiated version of *KidBiz3000®* did not meet WWC group design standards because the study (Tracey & Young, 2004) did not establish baseline equivalence for the intervention and comparison groups.

### Recommended Citation

What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education. (2018, February). *Adolescent Literacy intervention report: Achieve3000®*. Retrieved from <https://whatworks.ed.gov>

## WWC Rating Criteria

### Criteria used to determine the rating of a study

Study rating	Criteria
<b>Meets WWC group design standards without reservations</b>	A study that provides strong evidence for an intervention's effectiveness, such as a well-implemented RCT.
<b>Meets WWC group design standards with reservations</b>	A study that provides weaker evidence for an intervention's effectiveness, such as a QED or an RCT with high attrition that has established equivalence of the analytic samples.

### Criteria used to determine the rating of effectiveness for an intervention

Rating of effectiveness	Criteria
<b>Positive effects</b>	Two or more studies show statistically significant positive effects, at least one of which met WWC group design standards without reservations, AND No studies show statistically significant or substantively important negative effects.
<b>Potentially positive effects</b>	At least one study shows a statistically significant or substantively important positive effect, AND No studies show a statistically significant or substantively important negative effect AND fewer or the same number of studies show indeterminate effects than show statistically significant or substantively important positive effects.
<b>Mixed effects</b>	At least one study shows a statistically significant or substantively important positive effect AND at least one study shows a statistically significant or substantively important negative effect, but no more such studies than the number showing a statistically significant or substantively important positive effect, OR At least one study shows a statistically significant or substantively important effect AND more studies show an indeterminate effect than show a statistically significant or substantively important effect.
<b>Potentially negative effects</b>	One study shows a statistically significant or substantively important negative effect and no studies show a statistically significant or substantively important positive effect, OR Two or more studies show statistically significant or substantively important negative effects, at least one study shows a statistically significant or substantively important positive effect, and more studies show statistically significant or substantively important negative effects than show statistically significant or substantively important positive effects.
<b>Negative effects</b>	Two or more studies show statistically significant negative effects, at least one of which met WWC group design standards without reservations, AND No studies show statistically significant or substantively important positive effects.
<b>No discernible effects</b>	None of the studies shows a statistically significant or substantively important effect, either positive or negative.

### Criteria used to determine the extent of evidence for an intervention

Extent of evidence	Criteria
<b>Medium to large</b>	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.
<b>Small</b>	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 fewer than 350 students, AND, assuming 25 students in a class, a total of fewer than 14 classrooms across studies.

### Glossary of Terms

**Attrition** Attrition occurs when an outcome variable is not available for all subjects initially assigned to the intervention and comparison groups. If a randomized controlled trial (RCT) or regression discontinuity design (RDD) study has high levels of attrition, the validity of the study results can be called into question. An RCT with high attrition cannot receive the highest rating of *Meets WWC Group Design Standards without Reservations*, but can receive a rating of *Meets WWC Group Design Standards with Reservations* if it establishes baseline equivalence of the analytic sample. Similarly, the highest rating an RDD with high attrition can receive is *Meets WWC RDD Standards with Reservations*.

For single-case design research, attrition occurs when an individual fails to complete all required phases or data points in an experiment, or when the case is a group and individuals leave the group. If a single-case design does not meet minimum requirements for phases and data points within phases, the study cannot receive the highest rating of *Meets WWC Pilot Single-Case Design Standards without Reservations*.

**Baseline** A point in time before the intervention was implemented in group design research and in regression discontinuity design studies. When a study is required to satisfy the baseline equivalence requirement, it must be done with characteristics of the analytic sample at baseline. In a single-case design experiment, the baseline condition is a period during which participants are not receiving the intervention.

**Clustering adjustment** An adjustment to the statistical significance of a finding when the units of assignment and analysis differ. When random assignment is carried out at the cluster level, outcomes for individual units within the same clusters may be correlated. When the analysis is conducted at the individual level rather than the cluster level, there is a mismatch between the unit of assignment and the unit of analysis, and this correlation must be accounted for when assessing the statistical significance of an impact estimate. If the correlation is not accounted for in a mismatched analysis, the study may be too likely to report statistically significant findings. To fairly assess an intervention's effects, in cases where study authors have not corrected for the clustering, the WWC applies an adjustment for clustering when reporting statistical significance.

**Confounding factor** A confounding factor is a component of a study that is completely aligned with one of the study conditions, making it impossible to separate how much of the observed effect was due to the intervention and how much was due to the factor.

**Design** The method by which intervention and comparison groups are assigned (group design and regression discontinuity design) or the method by which an outcome measure is assessed repeatedly within and across different phases that are defined by the presence or absence of an intervention (single-case design). Designs eligible for WWC review are randomized controlled trials, quasi-experimental designs, regression discontinuity designs, and single-case designs.

**Effect size** The effect size is a measure of the magnitude of an effect. The WWC uses a standardized measure to facilitate comparisons across studies and outcomes.

**Eligibility** A study is eligible for review and inclusion in this report if it falls within the scope of the review protocol and uses either an experimental or matched comparison group design.

**Equivalence** A demonstration that the analytic sample groups are similar on observed characteristics defined in the review area protocol.

**Extent of evidence** An indication of how much evidence from group design studies supports the findings in an intervention report. The extent of evidence categorization for intervention reports focuses on the number and sizes of studies of the intervention in order to give an indication of how broadly findings may be applied to different settings. There are two extent of evidence categories: small and medium to large.

- **small:** includes only one study, or one school, or findings based on a total sample size of less than 350 students and 14 classrooms (assuming 25 students in a class)
- **medium to large:** includes more than one study, more than one school, and findings based on a total sample of at least 350 students or 14 classrooms

**Gain scores** The result of subtracting the pretest from the posttest for each individual in the sample. Some studies analyze gain scores instead of the unadjusted outcome measure as a method of accounting for the baseline measure when estimating the effect of an intervention. The WWC reviews and reports findings from analyses of gain scores, but gain scores do not satisfy the WWC's requirement for a statistical adjustment under the baseline equivalence requirement. This means that a study that must satisfy the baseline equivalence requirement and has baseline differences between 0.05 and 0.25 standard deviations *Does Not Meet WWC Group Design Standards* if the study's only adjustment for the baseline measure was in the construction of the gain score.

**Group design** A study design in which outcomes for a group receiving an intervention are compared to those for a group not receiving the intervention. Comparison group designs eligible for WWC review are randomized controlled trials and quasi-experimental designs.

**Improvement index** Along a percentile distribution of individuals, the improvement index represents the gain or loss of the average individual due to the intervention. As the average individual starts at the 50th percentile, the measure ranges from -50 to +50.

**Intervention** An educational program, product, practice, or policy aimed at improving student outcomes.

**Intervention report** A summary of the findings of the highest-quality research on a given program, product, practice, or policy in education. The WWC searches for all research studies on an intervention, reviews each against design standards, and summarizes the findings of those that meet WWC design standards.

**Multiple comparison adjustment** An adjustment to the statistical significance of results to account for multiple comparisons in a group design study. The WWC uses the Benjamini-Hochberg (BH) correction to adjust the statistical significance of results within an outcome domain when study authors perform multiple hypothesis tests without adjusting the  $p$ -value. The BH correction is used in three types of situations: studies that tested multiple outcome measures in the same outcome domain with a single comparison group; studies that tested a given outcome measure with multiple comparison groups; and studies that tested multiple outcome measures in the same outcome domain with multiple comparison groups. Because repeated tests of highly correlated constructs will lead to a greater likelihood of mistakenly concluding that the impact was different from zero, in all three situations, the WWC uses the BH correction to reduce the possibility of making this error. The WWC makes separate adjustments for primary and secondary findings.

**Outcome domain** A group of closely-related outcomes. A domain is the organizing construct for a set of related outcomes through which studies claim effectiveness.

<b>Quasi-experimental design (QED)</b>	A quasi-experimental design (QED) is a research design in which study participants are assigned to intervention and comparison groups through a process that is not random.
<b>Randomized controlled trial (RCT)</b>	A randomized controlled trial (RCT) is an experiment in which eligible study participants are randomly assigned to intervention and comparison groups.
<b>Rating of effectiveness</b>	For group design research, the WWC rates the effectiveness of an intervention in each domain based on the quality of the research design and the magnitude, statistical significance, and consistency in findings. For single-case design research, the WWC rates the effectiveness of an intervention in each domain based on the quality of the research design and the consistency of demonstrated effects. The criteria for the ratings of effectiveness are given in the WWC Rating Criteria on p. 20.
<b>Regression discontinuity design (RDD)</b>	A design in which groups are created using a continuous scoring rule. For example, students may be assigned to a summer school program if they score below a preset point on a standardized test, or schools may be awarded a grant based on their score on an application. A regression line or curve is estimated for the intervention group and similarly for the comparison group, and an effect occurs if there is a discontinuity in the two regression lines at the cutoff.
<b>Single-case design</b>	A research approach in which an outcome variable is measured repeatedly within and across different conditions that are defined by the presence or absence of an intervention.
<b>Standard deviation</b>	The standard deviation of a measure shows how much variation exists across observations in the sample. A low standard deviation indicates that the observations in the sample tend to be very close to the mean; a high standard deviation indicates that the observations in the sample tend to be spread out over a large range of values.
<b>Statistical significance</b>	Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups. The WWC labels a finding statistically significant if the likelihood that the difference is due to chance is less than 5% ( $p < .05$ ).
<b>Study rating</b>	The result of the WWC assessment of a study. The rating is based on the strength of the evidence of the effectiveness of the educational intervention. Studies are given a rating of <i>Meets WWC Design Standards without Reservations</i> , <i>Meets WWC Design Standards with Reservations</i> , or <i>Does Not Meet WWC Design Standards</i> , based on the assessment of the study against the appropriate design standards. The WWC has design standards for group design, single-case design, and regression discontinuity design studies.
<b>Substantively important</b>	A substantively important finding is one that has an effect size of 0.25 or greater, regardless of statistical significance.
<b>Systematic review</b>	A review of existing literature on a topic that is identified and reviewed using explicit methods. A WWC systematic review has five steps: 1) developing a review protocol; 2) searching the literature; 3) reviewing studies, including screening studies for eligibility, reviewing the methodological quality of each study, and reporting on high quality studies and their findings; 4) combining findings within and across studies; and, 5) summarizing the review.

Please see the [WWC Procedures and Standards Handbook \(version 3.0\)](#) for additional details.



An **intervention report** summarizes the findings of high-quality research on a given program, practice, or policy in education. The WWC searches for all research studies on an intervention, reviews each against evidence standards, and summarizes the findings of those that meet standards.

This intervention report was prepared for the WWC by Mathematica Policy Research under contract ED-IES-13-C-0010.