WWC Intervention Report

A summary of findings from a systematic review of the evidence

ies INSTITUTE OF EDUCATION SCIENCES

TAP™: The System for Teacher and Student Advancement

Report Contents

This intervention report presents findings from a systematic review of *TAP*™ conducted using the WWC Procedures and Standards Handbook, version 3.0, and the Teacher Training, Evaluation, and Compensation review protocol, version 3.1.

Program Description[1]

TAP™: *The System for Teacher and Student Advancement* (formerly known as the *Teacher Advancement Program*) is a comprehensive educator effectiveness program that aims to improve student achievement through supports and incentives that attract, retain, develop, and motivate effective teachers. The program provides teachers with leadership opportunities and associated salary increases; ongoing, school-based professional development; rigorous evaluations; and annual performance bonuses based on a combination of teacher value added to student achievement and observations of their classroom teaching.

Research[2]

The What Works Clearinghouse (WWC) identified one study of *TAP*™ that both falls within the scope of the Teacher Training, Evaluation, and Compensation topic area and meets WWC group design standards. This one study meets WWC group design standards with reservations. The study included 7,661 students in grades 4–8 in 34 Chicago elementary (grades K–8) schools.

The WWC considers the extent of evidence for having a teacher in a *TAP*™-implementing school on the academic achievement of students in grades 4–8 to be small for three student outcome domains—science achievement, English language arts achievement, and mathematics achievement. There were no studies that meet WWC design standards in the three other student outcome domains and the six teacher outcome domains, so this intervention report does not report on the effectiveness of *TAP*™ for those domains. (See the Effectiveness Summary on p. 6 for more details of effectiveness by domain.)

Effectiveness

Having a teacher in a *TAP*™-implementing school was found to have no discernible effects on science achievement for students in grades 4 and 7 or on English language arts and mathematics achievement for students in grades 4–8.

Table 1. Summary of findings[4,5]

Outcome domain	Rating of effectiveness	Improvement index *(percentile points)*		Number of studies	Number of students	Extent of evidence
		Average	Range			
Science achievement	No discernible effects	+5	na	1	1,717	Small
English language arts achievement	No discernible effects	0	na	1	7,661	Small
Mathematics achievement	No discernible effects	–1	na	1	7,656	Small

na = not applicable

Program Information

Background

Lowell Milken and educational experts at the Milken Family Foundation established *TAP*™ in 1999. The program is managed by the National Institute for Excellence in Teaching (NIET). The organization's address is 1250 Fourth Street, Santa Monica, CA 90401. Web: www.niet.org and www.tapsystem.org. Telephone: 310-570-4860.

Program details

NIET identifies clusters of potential *TAP*™ schools through its partnerships with districts, states, and universities. Identified schools that are interested in adopting *TAP*™ must submit an application, and NIET selects those applicant schools that demonstrate the capacity to implement *TAP*™, the ability to fund the system, and strong faculty support (that is, approved through a faculty vote). Once a school is selected to implement *TAP*™, leaders from those schools receive NIET training, materials, and tools to implement all four of the program's core elements. NIET also works with its *TAP*™ partner schools to obtain appropriate funding for services and to sustain *TAP*™'s core elements.

The four core elements of *TAP*™ are:

- ***Multiple career paths.*** Schools implementing *TAP*™ create new opportunities for high-performing teachers to take on additional leadership responsibilities by becoming mentor or master teachers. These individuals serve together with principals and assistant principals on a *TAP*™ Leadership Team, which is responsible for participating in trainings, analyzing student data, setting student learning goals and achievement plans, evaluating traditional teachers (called "career teachers"), and providing individual and team coaching. Master teachers also lead professional learning communities (called "cluster groups") and provide oversight to mentor teachers, who in turn provide additional support to career teachers. Mentor and master teachers receive release time and additional compensation to perform their leadership duties. *TAP*™ currently recommends annual salary augmentations of $5,000 to $8,000 for mentor teachers and $8,000 to $12,000 for master teachers, depending on local budgets.
- ***Ongoing applied professional growth.*** Teachers in *TAP*™ schools meet for 1 hour per week in grade- or subject-based groups led by mentor or master teachers. This collaborative mentoring and planning time is intended to help teachers learn research-based instructional strategies to meet the specific needs of their students. Mentor and master teachers also provide career teachers with individual, classroom-based support through activities such as demonstrating lessons, team-teaching, conducting observations, and providing feedback.
- ***Instructionally-focused accountability.*** Teachers in *TAP*™ schools are evaluated using three measures: classroom achievement growth, classroom observations, and school-wide achievement growth (or the latter two measures for teachers in non-tested subjects and grades). NIET trains and certifies members of the *TAP*™ Leadership Team to conduct observations using the *TAP*™ *Teaching Skills, Knowledge, and Responsibilities Performance Standards*. Each teacher is formally evaluated four times per year and receives feedback and coaching in a post-conference following each observation. Student achievement growth is assessed using teachers' grade- and subject-specific value-added scores derived by students' average growth trajectories per their state's standardized test.
- ***Performance-based compensation.*** NIET expects *TAP*™ schools to develop a performance award pool of about $2,000 to $3,000 per teacher for use as bonuses to the teachers who attain a minimum score on their evaluations. *TAP*™ schools also have the option to offer enhanced compensation to principals based on a locally-determined bonus structure. NIET can provide guidance on measures that can be used to determine a principal's eligibility for and amount of compensation (e.g., school-wide achievement growth, the *TAP*™ Leadership Team Rubric, and other valid and reliable 360° instruments).

Cost

For a typical school of 25 teachers and 600 students, the cost of *TAP*™ implementation is about $250 per student. Implementation costs include onsite and online support, the cost of a master teacher, salary stipends for other teacher leaders in the school, and bonuses. However, the cost per student may vary depending on the local cost of living, student/teacher ratios, and whether existing infrastructure can be leveraged (e.g., reading coaches who can become master teachers). Some districts use local or federal funds (Title I, Title II, or School Improvement Grants) to cover *TAP*™ costs. NIET can assist schools in identifying ways to leverage existing resources and local or federal funds to reduce the cost of *TAP*™.

Research Summary

The WWC identified nine eligible studies that investigated the effects of *TAP*™ teachers on academic achievement for students in grades 4–8. An additional 27 studies were identified but do not meet WWC eligibility criteria for review in this topic area. Citations for all 36 studies are in the References section, which begins on p. 8.

Table 2. Scope of reviewed research

Grade	4–8
Delivery method	Whole school
Program type	Teacher level

The WWC reviewed nine eligible studies against group design standards. One study (Glazerman & Seifullah, 2012) uses two designs to answer research questions: a cluster randomized controlled trial and a quasi-experimental design. Both designs meet WWC group design standards with reservations. The study is summarized in this report. Eight studies do not meet WWC group design standards.

Summary of studies meeting WWC group design standards without reservations

No studies of *TAP*™ met WWC group design standards without reservations.

Summary of study meeting WWC group design standards with reservations

Glazerman and Seifullah (2012) presented the implementation and impact findings resulting from Chicago Public Schools' phased roll-out of *TAP*™ (called "*Chicago TAP*") across four cohorts that included 34 randomly assigned elementary (grades K–8) schools.

The cluster randomized controlled trial component of the study featured two lotteries. One lottery randomly assigned an initial group of 16 recruited schools to either implement *TAP*™ immediately (Cohort 1: eight schools) or delay implementing *TAP*™ for 1 academic year (Cohort 2: eight schools). The second lottery, conducted 2 years later, randomly assigned another group of 18 schools to either implement *TAP*™ immediately (Cohort 3: nine schools) or delay implementation for 1 academic year (Cohort 4: nine schools). Thus, Cohort 2 served as a non-*TAP*™ comparison group for Cohort 1 during Cohort 1's first year of implementation, and Cohort 4 served as a non-*TAP*™ comparison group for Cohort 3 during Cohort 3's first year of implementation. Estimates of program impact after 1 year of implementation were then made by measuring the changes in student achievement for the *TAP*™ group of schools (Cohorts 1 and 3 pooled together) compared to the changes for both delayed implementation cohorts (Cohorts 2 and 4 pooled together). The authors also reported supplemental achievement findings from a quasi-experimental analysis that compared schools that implemented *TAP*™ (Cohorts 1 through 4) to a matched group of non-*TAP*™ schools.

Student achievement was assessed each March using scores on the science, reading, and mathematics sections of the Illinois Standards Achievement Test (ISAT), with the prior year's score being used as a pretest measure. The analytic samples for the cluster randomized controlled trial were: 1,717 grade 4 and 7 students (808 *TAP*™ and 909 comparison) for science achievement; 7,661 grade 4–8 students (3,717 *TAP*™ and 3,944 comparison) for reading achievement, which falls in the English language arts achievement domain; and 7,656 grade 4–8 students (3,714 *TAP*™ and 3,942 comparison) for mathematics achievement.[6]

Because the cluster randomized controlled trial analysis of the impact of *TAP*™ teachers on student achievement uses data from students who were present at the time of randomization of schools and those who joined the schools after randomization, the analysis is not eligible for the rating of *meets WWC group design standards without reservations*. The study demonstrated baseline equivalence of the analytic samples and, therefore, meets *WWC group design standards with reservations*.

Effectiveness Summary

The WWC review of *TAP*™ for the Teacher Training, Evaluation, and Compensation topic area includes both student and teacher outcomes. The review includes student outcomes in six domains: science achievement, English language arts achievement, mathematics achievement, social studies achievement, general achievement, and student progression. The review includes teacher outcomes in six domains: teacher instruction, teacher attendance, student growth scores, teacher retention at the school, teacher retention in the school district, and teacher retention in the profession. The one study of *TAP*™ that meets WWC group design standards reported findings in three of the six student-focused domains: (a) science achievement, (b) English language arts achievement, and (c) mathematics achievement.[7] The findings below present the authors' estimates and WWC-calculated estimates of the size and statistical significance of the effects of *TAP*™ teachers on students in grades 4–8. The supplemental findings based on the quasi-experimental analysis were similar in size and statistical significance to the findings from the cluster randomized controlled trial and are reported in the appendix. 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. 22.

Summary of effectiveness for the science achievement domain

One study that meets WWC group design standards with reservations reported findings in the science achievement domain.

Glazerman and Seifullah's (2012) analysis from the cluster randomized controlled trial examined one outcome in the science achievement domain: a score from the ISAT assessment for science. The authors found, and the WWC confirmed, that the difference between schools in spring of their first year of *TAP*™ implementation (Cohort 3) and schools that had not yet implemented *TAP*™ (Cohort 4) was not statistically significant.[8] According to WWC criteria, the effect size was not large enough to be considered substantively important (i.e., an effect size of at least 0.25). The WWC characterizes these study findings as an indeterminate effect.

Thus, for the science achievement domain, one study showed an indeterminate effect. This results in a rating of no discernible effects, with a small extent of evidence.

Table 3. Rating of effectiveness and extent of evidence for the science achievement domain

Rating of effectiveness	Criteria met
No discernible effects *No affirmative evidence of effects*	In the one study that reported findings, the estimated impact of the intervention on outcomes in the *science achievement* domain was neither statistically significant nor large enough to be substantively important.
Extent of evidence	**Criteria met**
Small	One study that included 1,717 students in 18 schools reported evidence of effectiveness in the *science achievement* domain.

Summary of effectiveness for the English language arts achievement domain

One study that meets WWC group design standards with reservations reported findings in the English language arts achievement domain.

Glazerman and Seifullah's (2012) analysis from the cluster randomized controlled trial examined one outcome in the English language arts achievement domain: a score from the ISAT assessment in reading. The authors found, and the WWC confirmed, that the difference between schools in spring of their first year of *TAP*™ implementation (Cohorts 1 and 3) and schools that had not yet implemented *TAP*™ (Cohorts 2 and 4) was not statistically significant. According to WWC criteria, the effect size was not large enough to be considered substantively important (i.e., an effect size of at least 0.25). The WWC characterizes these study findings as an indeterminate effect.

Thus, for the English language arts achievement domain, one study showed an indeterminate effect. This results in a rating of no discernible effects, with a small extent of evidence.

Table 4. Rating of effectiveness and extent of evidence for the English language arts achievement domain

Rating of effectiveness	Criteria met
No discernible effects *No affirmative evidence of effects.*	In the one study that reported findings, the estimated impact of the intervention on outcomes in the *English language arts achievement* domain was neither statistically significant nor large enough to be substantively important.
Extent of evidence	**Criteria met**
Small	One study that included 7,661 students in 34 schools reported evidence of effectiveness in the *English language arts achievement* domain.

Summary of effectiveness for the mathematics achievement domain

One study that meets WWC group design standards with reservations reported findings in the mathematics achievement domain.

Glazerman and Seifullah's (2012) analysis from the cluster randomized controlled trial examined one outcome in the mathematics achievement domain: a score from the ISAT assessment in mathematics. The authors found, and the WWC confirmed, that the difference between schools in spring of their first year of *TAP*™ implementation (Cohorts 1 and 3) and schools that had not yet implemented *TAP*™ (Cohorts 2 and 4) was not statistically significant. According to WWC criteria, the effect size was not large enough to be considered substantively important (i.e., an effect size of at least 0.25). The WWC characterizes these study findings as an indeterminate effect.

Thus, for the mathematics achievement domain, one study showed an indeterminate effect. This results in a rating of no discernible effects, with a small extent of evidence.

Table 5. Rating of effectiveness and extent of evidence for the mathematics achievement domain

Rating of effectiveness	Criteria met
No discernible effects *No affirmative evidence of effects.*	In the one study that reported findings, the estimated impact of the intervention on outcomes in the *mathematics achievement* domain was neither statistically significant nor large enough to be substantively important.
Extent of evidence	**Criteria met**
Small	One study that included 7,656 students in 34 schools reported evidence of effectiveness in the *mathematics achievement* domain.

References

Studies that meet WWC group design standards without reservations

None.

Study that meets WWC group design standards with reservations

Glazerman, S., & Seifullah, A. (2012). *An evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after four years. Final report*. Washington, DC: Mathematica Policy Research, Inc. http://files.eric.ed.gov/fulltext/ED530098.pdf.

Additional sources:

Glazerman, S., McKie, A., & Carey, N. (2009). *An evaluation of the Teacher Advancement Program (TAP) in Chicago: Year one impact report*. Washington, DC: Mathematica Policy Research, Inc. http://files.eric.ed.gov/fulltext/ED507502.pdf.

Glazerman, S., & Seifullah, A. (2010). *An evaluation of the Teacher Advancement Program (TAP) in Chicago: Year two impact report*. Washington, DC: Mathematica Policy Research, Inc. http://files.eric.ed.gov/fulltext/ED510712.pdf.

Studies that do not meet WWC group design standards

Eckert, J. (2013). South Carolina TAP. In *Increasing educator effectiveness: Lessons learned from Teacher Incentive Fund* sites (pp. 23–27, 50). Santa Monica, CA: National Institute for Excellence in Teaching. Retrieved from http://www.niet.org The study does not meet WWC evidence standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Additional source:

Eckert, J. (2010). South Carolina Department of Education and Florence County School District Three, SC – TAP. In *Performance-based compensation: Design and implementation at six Teacher Incentive Fund* sites (pp. 6, 25–28, 40). Chicago, IL and Seattle, WA: Joyce Foundation and Bill & Melinda Gates Foundation. Retrieved from http://www.tapsystem.org

Grant, G. M. (2010). *An investigation of the Teacher Advancement Program and student performance in urban schools* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3442745) The study does not meet WWC evidence standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Hudson, S. (2010). *The effects of performance-based teacher pay on student achievement* (SIEPR Discussion Paper 09-023). Stanford, CA: Stanford Institute for Economic Policy Research, Stanford University. The study does not meet WWC evidence standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Mann, D., & Leutscher, T. (2012). *Indiana's TIF/TAP program: Report of the year one evaluation*. Ashland, VA: Interactive Inc. Retrieved from http://ftp.goshenschools.org The study does not meet WWC evidence standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Additional source:

Eckert, J. (2013). Indiana Department of Education. In *Increasing educator effectiveness: Lessons learned from Teacher Incentive Fund sites* (pp. 33–35, 46). Santa Monica, CA: National Institute for Excellence in Teaching. Retrieved from http://www.niet.org

Mann, D., Leutscher, T., & Reardon, R. M. (2013). *Findings from a two-year examination of teacher engagement in TAP schools across Louisiana*. Ashland, VA: Interactive, Inc. Retrieved from http://www.niet.org The study does not meet WWC evidence standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Schacter, J., & Yeow, M. T. (2005). TAPping into high quality teachers: Preliminary results from the Teacher Advancement Program comprehensive school reform. *School Effectiveness & School Improvement*, *16*(3), 327–353. This study does not meet WWC evidence standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent

Additional sources:

Schacter, J., Schiff, T., Thum, Y. M., Fagnano, C., Bendotti, M., Solmon, L., … Milken, L. (2002). *The impact of the Teacher Advancement Program on student achievement, teacher attitudes, and job satisfaction*. Santa Monica, CA: Milken Family Foundation. Retrieved from http://www.tapsystem.org

Schacter, J., Thum, Y. M., Reifsneider, D., & Schiff, T. (2004). *The Teacher Advancement Program report two: Year three results from Arizona and year one results from South Carolina TAP schools*. Santa Monica, CA: Milken Family Foundation. Retrieved from http://www.tapsystem.org

Thum, Y. M. (2003). Measuring progress towards a goal: Estimating teacher productivity using a multivariate multilevel model for value-added analysis. *Sociological Methods & Research*, *32*(2), 153–207.

Solmon, L. C., White, J. T., Cohen, D., & Woo, D. (2007). *The effectiveness of the Teacher Advancement Program*. Santa Monica, CA: National Institute for Excellence in Teaching. The study does not meet WWC evidence standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Springer, M. G., Ballou, D., & Peng, A. (2014). Estimated effect of the Teacher Advancement Program on student test score gains. *Education Finance and Policy*, 9(2), 193–230. The study does not meet WWC evidence standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Additional source:

Springer, M. G., Ballou, D., & Peng, A. (2008). *Impact of the Teacher Advancement Program on student test score gains: Findings from an independent appraisal* (NCPI Working Paper 2008-19). Nashville, TN: National Center on Performance Incentives.

Studies that are ineligible for review using the Teacher Training, Evaluation, and Compensation Evidence Review Protocol

Almy, S., & Tooley, M. (2012). *Building and sustaining talent: Creating conditions in high-poverty schools that support effective teaching and learning*. Washington, DC: The Education Trust. http://files.eric.ed.gov/fulltext/ED543216.pdf. The study is ineligible for review because it does not use a comparison group design or a single-case design.

Additional source:

Eckert, J. (2013). Louisiana State Department of Education—TAP. In *Increasing educator effectiveness: Lessons learned from Teacher Incentive Fund sites* (pp. 41–43, 51). Santa Monica, CA: National Institute for Excellence in Teaching. Retrieved from http://www.niet.org

Center for High Impact Philanthropy. (2010). *High impact philanthropy to improve teaching quality in the U.S*. Philadelphia, PA: University of Pennsylvania, School of Social Policy & Practice. Retrieved from http://www.impact.upenn.edu The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Culbertson, J. (2012). Putting the value in teacher evaluation. *Phi Delta Kappan, 94*(3), 14–18. The study is ineligible for review because it does not use a comparison group design or a single-case design.

Daley, G., & Kim, L. (2010). *A teacher evaluation system that works* (NIET Working Paper). Santa Monica, CA: National Institute for Excellence in Teaching. Retrieved from http://files.eric.ed.gov/fulltext/ED533380.pdf The study is ineligible for review because it does not use a comparison group design or a single-case design.

Additional source:

Daley, G., & Kim, L. (2010). *A teacher evaluation system that works* (NIET Research Brief). Santa Monica, CA: National Institute for Excellence in Teaching. Retrieved from http://www.tapsystem.org

DeMonte, J. (2013). *High-quality professional development for teachers: Supporting teacher training to improve student learning*. Washington, DC: Center for American Progress. Retrieved from http://www.americanprogress.org The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Dispenzieri, M. M. (2009). *Teacher perception of the impact of the Teacher Advancement Program on student achievement* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3342481) The study is ineligible for review because it does not use a comparison group design or a single-case design.

Eckert, J. (2009). *More than widgets: TAP: A systemic approach to increased teaching effectiveness*. Santa Monica, CA: National Institute for Excellence in Teaching. Retrieved from http://www.tapsystem.org The study is ineligible for review because it does not use a comparison group design or a single-case design.

Eckert, J. (2010). School District of Philadelphia, PA – Charter Schools (Philly TAP). In *Performance-based compensation: Design and implementation at six Teacher Incentive Fund sites* (pp. 6, 21–24). Chicago, IL and Seattle, WA: Joyce Foundation and Bill & Melinda Gates Foundation. Retrieved from http://www.tapsystem.org The study is ineligible for review because it does not use a comparison group design or a single-case design.

Eckert, J. (2010). University of Texas System (Texas TAP). In *Performance-based compensation: Design and implementation at six Teacher Incentive Fund sites* (pp. 7, 29–32). Chicago, IL and Seattle, WA: Joyce Foundation and Bill & Melinda Gates Foundation. Retrieved from http://www.tapsystem.org The study is ineligible for review because it does not use a comparison group design or a single-case design.

Eckert, J. (2013). Algiers, Louisiana: NIET – TAP & Consortium of Algiers Charter Schools. In *Increasing educator effectiveness: Lessons learned from Teacher Incentive Fund sites* (pp. 5–9, 46–47). Santa Monica, CA: National Institute for Excellence in Teaching. Retrieved from http://www.niet.org The study is ineligible for review because it does not examine an intervention implemented in a way that falls within the scope of the review—the intervention is bundled with other components.

Additional sources:

Algiers Charter School Association. (2009). *TAP weaves a tapestry of achievement at Algiers Charter Schools*. New Orleans, LA: Author. Retrieved from http://www.tapsystem.org

Algiers Charter School Association. (2011). *Annual report 2011*. New Orleans, LA: Author. Retrieved from http://www.tapsystem.org

Eckert, J. (2010). National Institute for Excellence in Teaching – TAP: The System for Teacher and Student Advancement, Consortium of Algiers Charter Schools, New Orleans, LA. In *Performance-based compensation: Design and implementation at six Teacher Incentive Fund sites* (pp. 5, 8–12, 35–36). Chicago, IL and Seattle, WA: Joyce Foundation and Bill & Melinda Gates Foundation. Retrieved from http://www.tapsystem.org

Eckert, J. (2013). Knox County, Tennessee: NIET – TAP & Knox County Schools. In *Increasing educator effectiveness: Lessons learned from Teacher Incentive Fund sites* (pp. 37–39). Santa Monica, CA: National Institute for Excellence in Teaching. Retrieved from http://www.niet.org The study is ineligible for review because it does not use a comparison group design or a single-case design.

Fain, A. (2012). *Analyzing the South Carolina Teacher Advancement Program's effectiveness and its impact on teachers' professional growth* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3541515) The study is ineligible for review because it does not use a comparison group design or a single-case design.

Hebert, S. H. (2014). *Charter and non-charter schools: Show us the money* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3617170) The study is ineligible for review because it does not use a comparison group design or a single-case design.

Jerald, C. (2009). *Aligned by design: How teacher compensation reform can support and reinforce other educational reforms*. Washington, DC: Center for American Progress. Retrieved from http://www.americanprogress.org The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Jerald, C. D. (2012). *Movin' it and improvin' it!: Using both education strategies to increase teaching effectiveness*. Washington, DC: Center for American Progress. Retrieved from http://files.eric.ed.gov/fulltext/ED535645.pdf The study is ineligible for review because it does not use a comparison group design or a single-case design.

Jerald, C. D., & Van Hook, K. (2011). *More than measurement: The TAP system's lessons learned for designing better teacher evaluation systems*. Santa Monica, CA: National Institute for Excellence in Teaching. Retrieved from http://files.eric.ed.gov/fulltext/ED533382.pdf The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Natale, C. F., Bassett, K., Gaddis, L., & McKnight, K. (2013). *Creating sustainable teacher career pathways: A 21st century imperative*. Washington, DC and New York, NY: The National Network of State Teachers of the Year and the Center for Educator Effectiveness at Pearson. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

National Institute for Excellence in Teaching. (2010). *Voices from the field: Teachers describe their experience with a bold system of reform*. Santa Monica, CA: Author. Retrieved from http://www.tapsystem.org The study is ineligible for review because it does not use a comparison group design or a single-case design.

National Institute for Excellence in Teaching. (2012). *Beyond job-embedded: Ensuring that good professional development gets results*. Santa Monica, CA: Author. Retrieved from http://www.niet.org This study is ineligible for review because it does not include an outcome within a domain specified in the protocol.

National Institute for Excellence in Teaching. (2012). *The effectiveness of TAP: Research summary 2012*. Santa Monica, CA: Author. Retrieved from http://www.tapsystem.org The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Paulmann, G. (2009). *Master teachers' critical practice and student learning strategies: A case study in an urban school district* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3393071) The study is ineligible for review because it does not use a comparison group design or a single-case design.

Pieczura, M. (2012). Weighing the pros and cons of TAP. *Educational Leadership, 70*(3), 70–72. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Portie, J. (2009). *Investigation of pre-service support for student teachers* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3447736) The study is ineligible for review because it does not include an outcome within a domain specified in the protocol.

Ritter, G. W., & Barnett, J. H. (2013). *A straightforward guide to teacher merit pay: Encouraging and rewarding schoolwide improvement*. Chicago, IL: Corwin Press. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

The Teaching Commission. (2006). *Teaching at risk: Progress & potholes*. New York, NY: Author. Retrieved from http://www.nctq.org The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

White, J. T. (2006). *Does principal leadership really matter? An analysis of the relationship between implementation of the Teacher Advancement Program and student achievement* (Doctoral dissertation). Available from ProQuest

Dissertations and Theses database. (UMI No. 3216247) The study is ineligible for review because it does not use a comparison group design or a single-case design.

Womack, D. (2011). *A case study of middle school teachers' perceptions of their professional development experiences with TAP—The System for Teacher and Student Advancement* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3501087) The study is ineligible for review because it does not use a comparison group design or a single-case design.

Appendix A: Research details for Glazerman & Seifullah (2012)[9]

Glazerman, S., & Seifullah, A. (2012). *An evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after four years.* Final report. Washington, DC: Mathematica Policy Research. http://files.eric.ed.gov/fulltext/ED530098.pdf.

Table A. Summary of findings

Meets WWC group design standards with reservations

Outcome domain	Sample size	Study findings	
		Average improvement index *(percentile points)*	Statistically significant
Science achievement	18 schools/1,717 students	+5	No
English language arts achievement	34 schools/7,661 students	0	No
Mathematics achievement	34 schools/7,656 students	–1	No

Setting

The study was conducted in Chicago Public Schools starting in the 2007–08 school year and continuing through the 2010–11 school year.

Study sample

Cluster Randomized Controlled Trial

A total of 34 public elementary (grades K–8) schools in Chicago participated in the cluster randomized controlled trial part of the study. More than 90% of the students in these schools were African American, and more than 95% were eligible for free or reduced-price lunch.

In spring 2007, 16 elementary schools were randomly assigned to begin *TAP*™ either in fall 2007 (eight schools in the *TAP*™ group [Cohort 1]) or in fall 2008 (eight schools in the comparison group [Cohort 2]). In spring 2009, 18 additional elementary schools were randomly assigned to begin *TAP*™ either in fall 2009 (nine schools in the *TAP*™ group [Cohort 3]) or in fall 2010 (nine schools in the comparison group [Cohort 4]).

Students in grades 4–8 were included in the analysis of the impact of *TAP*™ teachers on student achievement for the first year of *TAP*™ implementation: 1,717 students in the science achievement sample (808 *TAP*™ students and 909 comparison students), which is smaller than the others because standardized test data in science were available only for students in grades 4 and 7 and only for Cohorts 3 and 4; 7,661 students in the English language arts achievement sample (3,717 *TAP*™ students and 3,944 comparison students); and 7,656 students in the mathematics achievement sample (3,714 *TAP*™ students and 3,942 comparison students).[10]

Quasi-Experiment

The quasi-experimental portion of the study included six purposively selected *TAP*™ schools in addition to the 34 randomly assigned *TAP*™ schools. For the quasi-experiment, *TAP*™ schools from all four cohorts were matched to other schools in the district that were not participating in *TAP*™ on measures such as school size, teacher retention, student race/ethnicity, student achievement, student poverty, student special education status, student language proficiency, and charter school status. The authors used a propensity score matching procedure where *TAP*™ schools were matched to their nearest five neighbors, with replacement. The resulting sample consisted of students in about 40 *TAP*™ schools and about 100 non-*TAP*™ comparison schools.[11]

The analytic samples for the quasi-experimental analysis of the impact of *TAP*™ teachers on student achievement for the first year of *TAP*™ implementation were: 12,998 grade 4 and 7 students (2,464 *TAP*™ and 10,534 comparison) for science achievement; and 41,580 grade 4–8 students (8,097 *TAP*™ and 33,483 comparison) for both English language arts achievement and mathematics achievement.[10] The results from this non-experimental analysis are presented as supplemental findings in the appendix. The supplemental findings do not factor into the intervention's rating of effectiveness.

Intervention group

Under *TAP*™, teachers can earn extra pay and responsibilities by being promoted to mentor or master teachers and can earn annual performance bonuses based on a combination of their value added to student achievement and observations of their classroom teaching. Unlike with the national *TAP*™ model, the program as implemented in Chicago Public Schools (called *"Chicago TAP"*) did not measure value-added performance at the individual teacher (or classroom) level; rather, value added was measured at the school level in 2007–08 and 2008–09 and at both the school- and school-grade team levels in 2009–10 and 2010–11. In Chicago, *TAP*™ included weekly meetings of teachers and mentors, regular classroom observations by a school leadership team, and pay for principals who meet implementation benchmarks. In the first year of implementation, teachers in Cohorts 1, 2, and 3 (i.e., those implementing *TAP*™ in 2007–08 through 2009–10) received an average bonus of $1,100; teachers in Cohort 4 received an average bonus of $1,400 in 2010–11. Average bonuses increased to approximately $2,500 in the second and third years of implementation, and were $1,900 in the fourth year of implementation. Teachers and mentors met weekly, and mentors received an additional $7,000 per year. Master teachers (called "lead teachers" in Chicago) received $15,000.

Comparison group

For the cluster randomized controlled trial portion of the study, comparison schools were in a "business-as-usual" condition for a year and subsequently participated in *TAP*™. For the quasi-experimental portion of the study, comparison schools were in a "business-as-usual" condition and did not receive *TAP*™.

Outcomes and measurement

Student standardized test data on science (grades 4 and 7), English language arts (grades 4–8), and mathematics (grades 4–8) were obtained from Chicago Public Schools. For a more detailed description of these outcome measures, see Appendix B.

The study also analyzed teacher retention at the school, teacher retention in the school district, and teacher attitudes. However, the teacher retention outcomes are rated *does not meet WWC group design standards* because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated. Teacher attitudes were not included in this review because the outcomes fall outside of the domains of interest listed in the Teacher Training, Evaluation, and Compensation review protocol (version 3.1).

Support for implementation

The *TAP*™ model provides for observations of teachers by the principal, mentor teachers, and master teachers, all of whom undergo training and certification in using the Skills, Knowledge, and Responsibilities (SKR) rubric. SKR scores are based on observed classroom performance in four domains: designing and planning instruction, learning environment, instruction, and responsibilities.

Appendix B: Outcome measures for each domain

Science achievement	
Illinois Standards Achievement Test (ISAT): Science Assessment	The ISAT Science Assessment is a standardized statewide test administered to students in grades 4 and 7. Assessment scores were obtained from Chicago Public Schools (CPS) (as cited in Glazerman & Seifullah, 2012).
English language arts achievement	
ISAT: Reading Assessment	The ISAT Reading Assessment is a standardized statewide test administered to students in grades 3–8. Assessment scores were obtained from CPS (as cited in Glazerman & Seifullah, 2012).
Mathematics achievement	
ISAT: Mathematics Assessment	The ISAT Mathematics Assessment is a standardized statewide test administered to students in grades 3–8. Assessment scores were obtained from CPS (as cited in Glazerman & Seifullah, 2012).

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

Outcome measure	Study sample	Sample size	Mean (standard deviation) Intervention group	Mean (standard deviation) Comparison group	WWC calculations Mean difference	WWC calculations Effect size	WWC calculations Improvement index	*p*-value
Glazerman & Seifullah (2012)[a]								
ISAT: Science Assessment (cluster randomized controlled trial)	Grade 4 and 7 students after 1 year of *TAP*™	1,717 students	204.3 (31.0)	200.6 (31.0)	3.7	0.12	+5	.12
Domain average for science achievement (Glazerman & Seifullah, 2012)						**0.12**	**+5**	**Not statistically significant**
Domain average for science achievement across all studies						**0.12**	**+5**	**na**

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. ISAT = Illinois Standards Achievement Test.

[a] For Glazerman and Seifullah (2012), no corrections for clustering or multiple comparisons and no difference-in-differences adjustments were needed. The *p*-value presented here was reported in the original study. The standard deviations were provided by the study authors at the WWC's request. The authors also provided a corrected comparison group mean. This study is characterized as having an indeterminate effect because the estimated effect is neither statistically significant nor 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 English language arts achievement domain

Outcome measure	Study sample	Sample size	Mean (standard deviation) Intervention group	Mean (standard deviation) Comparison group	WWC calculations Mean difference	WWC calculations Effect size	WWC calculations Improvement index	*p*-value
Glazerman & Seifullah (2012)[a]								
ISAT: Reading Assessment (cluster randomized controlled trial)	Grade 4–8 students after 1 year of *TAP*™	7,661 students	221.3 (26.5)	221.0 (27.0)	0.3	0.01	0	> .10
Domain average for English language arts achievement (Glazerman & Seifullah, 2012)						**0.01**	**0**	**Not statistically significant**
Domain average for English language arts achievement across all studies						**0.01**	**0**	**na**

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. ISAT = Illinois Standards Achievement Test.

[a] For Glazerman and Seifullah (2012), no corrections for clustering or multiple comparisons and no difference-in-differences adjustments were needed. The *p*-value presented here was reported in the original study. The standard deviations were provided by the study authors at the WWC's request. This study is characterized as having an indeterminate effect because the estimated effect is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

Appendix C.3: Findings included in the rating for the mathematics 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	
Glazerman & Seifullah (2012)[a]								
ISAT: Mathematics Assessment (cluster randomized controlled trial)	Grade 4–8 students after 1 year of *TAP*™	7,656 students	233.4 (25.1)	234.3 (28.8)	–0.9	–0.03	–1	> .10
Domain average for mathematics achievement (Glazerman & Seifullah, 2012)						**–0.03**	**–1**	**Not statistically significant**
Domain average for mathematics achievement across all studies						**–0.03**	**–1**	**na**

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. ISAT = Illinois Standards Achievement Test.

[a] For Glazerman and Seifullah (2012), no corrections for clustering or multiple comparisons and no difference-in-differences adjustments were needed. The *p*-value presented here was reported in the original study. The standard deviations were provided by the study authors at the WWC's request. This study is characterized as having an indeterminate effect because the estimated effect is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

Appendix D1: Supplemental quasi-experimental design findings for the science 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	
Glazerman & Seifullah (2012)[a]								
ISAT: Science Assessment (quasi-experimental design)	Grade 4 and 7 students after 1 year of *TAP*™	12,998 students	205.0 (28.4)	203.7 (27.9)	1.3	0.05	+2	> .10

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. ISAT = Illinois Standards Achievement Test.

[a] For Glazerman and Seifullah (2012), no corrections for clustering or multiple comparisons and no difference-in-differences adjustments were needed. The *p*-value presented here was reported in the original study. The WWC calculated the intervention group mean by adding the impact of the intervention (the estimated coefficient on the intervention group indicator from a regression model) to the unadjusted comparison group posttest mean. The analytic sample sizes, unadjusted means, and unadjusted standard deviations were provided by the study authors at the WWC's request.

Appendix D.2: Supplemental quasi-experimental design findings for the English language arts 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	
Glazerman & Seifullah (2012)[a]								
ISAT: Reading Assessment (quasi-experimental design)	Grade 4–8 students after 1 year of *TAP*™	41,580 students	222.2 (26.6)	222.4 (26.5)	–0.2	–0.01	0	> .10

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. ISAT = Illinois Standards Achievement Test.

[a] For Glazerman and Seifullah (2012), no corrections for clustering or multiple comparisons and no difference-in-differences adjustments were needed. The *p*-value presented here was reported in the original study. The WWC calculated the intervention group mean by adding the impact of the intervention (the estimated coefficient on the intervention group indicator from a regression model) to the unadjusted comparison group posttest mean. The analytic sample sizes, unadjusted means, and unadjusted standard deviations were provided by the study authors at the WWC's request.

Appendix D.3: Supplemental quasi-experimental design findings for the mathematics 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	
Glazerman & Seifullah (2012)[a]								
ISAT: Mathematics Assessment (quasi-experimental design)	Grade 4–8 students after 1 year of *TAP*™	41,580 students	235.9 (29.9)	235.5 (29.1)	0.4	0.01	+1	> .10

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. ISAT = Illinois Standards Achievement Test.

[a] For Glazerman and Seifullah (2012), no corrections for clustering or multiple comparisons and no difference-in-differences adjustments were needed. The *p*-value presented here was reported in the original study. The WWC calculated the intervention group mean by adding the impact of the intervention (the estimated coefficient on the intervention group indicator from a regression model) to the unadjusted comparison group posttest mean. The analytic sample sizes, unadjusted means, and unadjusted standard deviations were provided by the study authors at the WWC's request.

Endnotes

[1] The descriptive information for this program was obtained from publicly available sources: the program's websites (www.niet.org and www.tapsystem.org, downloaded February 2014). The WWC requests developers review the program description sections for accuracy from their perspective. The program description was provided to the developer in February 2014, and the WWC incorporated feedback from the developer. Further verification of the accuracy of the descriptive information for this program is beyond the scope of this review.

[2] The literature search reflects documents publicly available by July 2014. A single study review of Glazerman & Seifullah (2012) was released in February 2013, which rated both the cluster randomized controlled trial analysis of student achievement and the quasi-experimental analysis of teacher retention as *meets WWC group design standards with reservations*. However, the rating of the teacher retention analysis differs in this report due to the use of a different review protocol. The single study review protocol (version 2.0) under which the single study review was conducted requires quasi-experimental analyses to demonstrate equivalence on baseline measures of the outcome. The single study review concluded that baseline equivalence was demonstrated for the teacher retention analysis based on an examination of baseline measures of school-level teacher retention. The Teacher Training, Evaluation, and Compensation review protocol (version 3.1) under which the review for this intervention report was conducted requires quasi-experimental analyses of teacher retention in schools to demonstrate equivalence on baseline measures of (1) teacher experience, (2) student academic performance, (3) student race/ethnicity or a degree of disadvantage, and (4) a school-level measure of teacher retention. Because the authors could not provide these baseline measures for the analytic sample, the review for this intervention report concluded that baseline equivalence was not demonstrated for the teacher retention analysis. Therefore, the analysis of teacher retention *does not meet WWC group design standards*. The studies in this report were reviewed using the Standards from the WWC Procedures and Standards Handbook (version 3.0), along with those described in the Teacher Training, Evaluation, and Compensation review protocol (version 3.1). The evidence presented in this report is based on available research. Findings and conclusions may change as new research becomes available.

[3] Absence of conflict of interest: This intervention report includes a study conducted by staff from Mathematica Policy Research. Because Mathematica Policy Research is one of the contractors that administers the WWC, the study was reviewed by staff members from a different organization, who also prepared the intervention report. The report was then reviewed by the lead methodologist, a WWC Quality Assurance reviewer, and an external peer reviewer.

[4] For criteria used in the determination of the rating of effectiveness and extent of evidence, see the WWC Rating Criteria on p. 22. These improvement index numbers show the average and range of individual-level improvement indices for all findings across the studies.

[5] The following domains were not examined by studies that meet WWC design standards: social studies achievement, general achievement, student progression, teacher instruction, teacher attendance, student growth scores, and teacher retention in the profession. The one study that met standards examined outcomes in the teacher retention at the school and teacher retention in the school district domains; however, the outcomes are rated *do not meet WWC group design standards* because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

[6] The student analytic sample sizes by condition were provided by the study authors at the WWC's request.

[7] The one study that meets WWC group design standards—Glazerman & Seifullah (2012)—reported findings for outcomes in the teacher retention at the school and teacher retention at the school district domains; however, the outcomes are rated *do not meet WWC group design standards* because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

[8] The authors present these science achievement findings for only Cohorts 3 and 4, because science scores were not available during the first year of implementation for Cohort 1.

[9] The WWC identified two additional sources related to Glazerman & Seifullah (2012). These studies do not contribute unique information to Appendix A and are not listed here.

[10] The student analytic sample sizes by condition were provided by the study authors at the WWC's request.

[11] The number of *TAP*™ schools in the quasi-experimental analysis of the impact on student achievement for the first year of *TAP*™ implementation differed from the number of *TAP*™ schools in the analogous cluster randomized controlled trial analysis for two reasons. First, in addition to the 34 randomly assigned *TAP*™ schools, the quasi-experimental analytic sample included two purposively assigned charter schools and four "replacement schools" that were selected for *TAP*™ when other schools closed or discontinued the program. Second, whereas the cluster randomized controlled trial analysis included all randomly assigned *TAP*™ schools, even if they discontinued the program, the quasi-experimental analysis dropped schools that exited the program or closed, along with their matched comparison schools, beginning in the school year the changes went into effect. The authors do not report the number of schools included specifically in the quasi-experimental analysis of student achievement after the first year of implementation; however, a table pertaining to the quasi-experimental study more broadly suggests that there were 39 *TAP*™ schools and 99 matched comparison schools (see notes for Table II.2 on p. 17).

Recommended Citation

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2015, July). *Teacher Training, Evaluation, and Compensation intervention report: TAP™: The System for Teacher and Student Advancement*. Retrieved from http://whatworks.ed.gov

WWC Rating Criteria

Criteria used to determine the rating of a study

Study rating	Criteria
Meets WWC group design standards without reservations	A study that provides strong evidence for an intervention's effectiveness, such as a well-implemented RCT.
Meets WWC group design standards with reservations	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
Positive effects	Two or more studies show statistically significant positive effects, at least one of which met WWC group design standards for a strong design, AND No studies show statistically significant or substantively important negative effects..
Potentially positive effects	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.
Mixed effects	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.
Potentially negative effects	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.
Negative effects	Two or more studies show statistically significant negative effects, at least one of which met WWC group design standards for a strong design, AND No studies show statistically significant or substantively important positive effects..
No discernible effects	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
Medium to large	The domain includes more than one study, AND The domain includes more than one school, AND The domain findings are based on a total sample size of at least 350 students, OR, assuming 25 students in a class, a total of at least 14 classrooms across studies.
Small	The domain includes only one study, OR The domain includes only one school, OR The domain findings are based on a total sample size of 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 participants initially assigned to the intervention and comparison groups. The WWC considers the total attrition rate and the difference in attrition rates across groups within a study.

Clustering adjustment If intervention assignment is made at a cluster level and the analysis is conducted at the student level, the WWC will adjust the statistical significance to account for this mismatch, if necessary.

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 design of a study is the method by which intervention and comparison groups were assigned.

Domain A domain is a group of closely related outcomes.

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 analysis sample groups are similar on observed characteristics defined in the review area protocol.

Extent of evidence An indication of how much evidence supports the findings. The criteria for the extent of evidence levels are given in the WWC Rating Criteria on p. 22.

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 When a study includes multiple outcomes or comparison groups, the WWC will adjust the statistical significance to account for the multiple comparisons, if necessary.

Quasi-experimental design (QED) 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.

Randomized controlled trial (RCT) A randomized controlled trial (RCT) is an experiment in which eligible study participants are randomly assigned to intervention and comparison groups.

Rating of effectiveness The WWC rates the effects of an intervention in each domain based on the quality of the research design and the magnitude, statistical significance, and consistency in findings. The criteria for the ratings of effectiveness are given in the WWC Rating Criteria on p. 22.

Single-case design 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.

Glossary of Terms

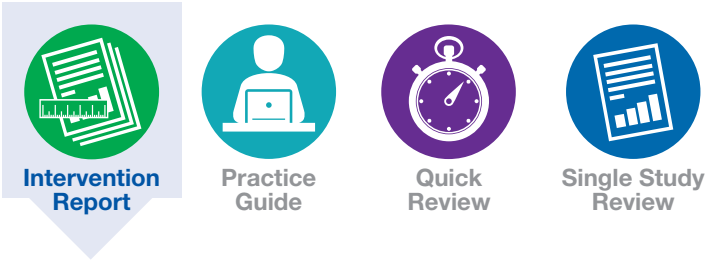
Standard deviation The standard deviation of a measure shows how much variation exists across observations in the sample. A low standard deviation indicates that the observations in the sample tend to be very close to the mean; a high standard deviation indicates that the observations in the sample tend to be spread out over a large range of values.

Statistical significance 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% ($p < .05$).

Substantively important A substantively important finding is one that has an effect size of 0.25 or greater, regardless of statistical significance.

Systematic review 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.