WWC Single Study Review

INSTITUTE OF EDUCATION SCIENCES

March 2015

WWC Review of the Report "Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment"[1,2]

The findings from this review do not reflect the full body of research evidence on teacher incentive programs.

What is this study about?

The study authors examined the impact of an intervention, known to participants as the *Talent Transfer Initiative (TTI)*, on average standardized test scores in mathematics and reading for over 14,000 students in 10 school districts. The *TTI* program enabled principals of low-performing schools to provide bonuses to high-performing teachers when they transferred to and stayed in the low-performing schools.

The study design is based on a randomized controlled trial that assigns "teacher teams" to potentially participate in the *TTI* program or to not participate in the *TTI* program. Prior to randomization, the study authors identified schools that had a teacher vacancy in a specified grade level and subject. A teacher team consisted of all teaching positions within the same grade level and subject. Schools were then matched based on the grade level and subject of the teacher teams. Within each matched pair of schools, the study authors randomly assigned the teacher teams to one of two conditions over the course of 2 years: either the principal was allowed to hire a high-performing teacher through the *TTI* program ("*TTI* team") to fill the vacancy in the teacher team, or the principal could hire a teacher to fill the vacancy through normal approaches ("comparison team").

The study authors measured mathematics and reading achievement for the students taught by the teacher teams in all 10 of the participating school districts (Cohorts 1 and 2) after the end of the first *TTI* program year and for a different wave of students taught by the teacher teams in seven of the 10 districts (Cohort 1 only) where *TTI* had been implemented for two program years following random assignment. Impacts were measured separately for teacher teams in elementary schools and middle schools.[3,4]

WWC Rating

The research described in this report meets WWC group design standards without reservations

The study is a well-executed clustered randomized controlled trial. **A subset of the analyses described in the study meets WWC group design standards without reservations.** Specifically, this rating pertains to all elementary school sample analyses, the first program year mathematics and reading middle school analyses, and the second program year reading middle school sample analyses. In all of these analyses, the study has low team-level attrition.

The second year mathematics achievement analysis for the middle school sample meets WWC group design standards with reservations because there was high team-level attrition, and the study demonstrated equivalence of the analytic intervention and comparison groups.[4]

The study authors acknowledge that all impact estimates, which are presented at the team-level, represent a combination of the intervention's effect on student achievement and the influence of the compositional changes, such as student-level mobility (for example, high-achieving students moving into *TTI* teacher teams) or teacher mobility (for example, principals moving higher- or lower-achieving teachers into participating teacher teams).

What did the study find?

The study authors reported, and the WWC confirmed, that elementary students from *TTI* teams scored higher, by a statistically significant margin, on mathematics and reading standardized tests at the end of the second program year than those from comparison teams. The authors found, and the WWC confirmed, no statistically significant impacts in test scores at the end of the first program year for both the elementary and middle school samples and no significant impacts at the end of program Year 2 for the middle school samples. None of the nonsignificant effects were large enough to be considered substantively important by the WWC.

Features of the *Talent Transfer Initiative (TTI)*

The *TTI* program offered principals the opportunity to fill vacant positions in schools that serve the lowest-performing students in the district with highly effective teachers. Eligible teachers who filled these vacant positions could receive up to $20,000 in bonuses over 2 academic years. Teachers were considered highly effective and thus eligible to volunteer to participate if they taught a class in grades 3–8 and were in the top 20% in their district on student test score growth. Bonuses were paid out in five installments over a 2-year period and were contingent upon both transferring into and remaining in the targeted school.

Appendix A: Study details

Glazerman, S., Protik, A., Teh, B., Bruch, J., & Max, J. (2013). *Transfer incentives for high-performing teachers: Final results from a multisite randomized experiment* (NCEE 2014-4003). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. http://files.eric.ed.gov/fulltext/ED544269.pdf.

Setting

The study was conducted in 114 elementary and middle schools in 10 large, diverse school districts in seven states. Schools served low-achieving, primarily minority students, with approximately 75% of students' families eligible for federal free or reduced-price lunch.

Study sample

The study design is based on a randomized controlled trial that assigns "teacher teams" to an intervention group or a comparison group. Prior to randomization, the study authors identified teacher vacancies in schools by grade level and subject. A teacher team consisted of all teaching positions within the same grade level and subject. Researchers matched teams across schools within a district based on subject, grade, and, when possible, other school characteristics such as student achievement ranking and student free and reduced-price lunch eligibility rates. Within each school district, the matched teams were then randomly assigned to the intervention and comparison conditions.

Randomization in seven school districts (Cohort 1) took place in the first year of the study, 2009–10, and teams in three additional school districts (Cohort 2) were randomized in the second year, 2010–11. In total, 165 teacher teams that had at least one vacancy were randomly assigned either to the *TTI* condition (85 teacher teams; 64 in Cohort 1 and 21 in Cohort 2) or to the comparison condition (80 teacher teams; 60 in Cohort 1 and 20 in Cohort 2). There were 92 teacher vacancies within the *TTI* teams and 88 vacancies within the comparison teams. The program Year 1 analyses focused on achievement of the students taught by teacher teams in all 10 Cohort 1 and Cohort 2 districts, and the program Year 2 analysis focused on a separate group of students taught by teacher teams from the second year of operation in the Cohort 1 districts. Specifically, the program Year 1 elementary school analyses included 97 teacher teams (49 *TTI* and 48 comparison), and the program Year 2 analyses included 90 teams (46 *TTI* and 44 comparison). The program Year 1 middle school analytic samples included up to 31 teams (16 *TTI* and 14–15 comparison teams, depending on the analysis), and the program Year 2 analysis included up to 20 teams (up to 10 in each of the *TTI* and comparison teams). Student-level sample sizes ranged from 2,355 to 8,038, depending on the analysis. Elementary school analyses focused on teacher teams in grades 3–5, and middle school analyses focused on teacher teams in grades 6–8. In some cases, teachers were included in analyses across different grade levels, and students were featured across different subject areas.

Intervention group

Principals were given the opportunity to fill any vacant teaching positions within a teacher team using *TTI* funds, which were intended to attract highly effective teachers to transfer to these schools. In total, 88% of vacancies in teams assigned to the *TTI* condition were filled using *TTI*, with 8% filled outside *TTI* and 4% remaining unfilled. Students in the *TTI* condition could have had a teacher who was offered *TTI* funds ("focal teachers") or another teacher within the *TTI* teacher team who had already been teaching at the school or who was hired through other means ("nonfocal teacher").[5]

Comparison group

Principals were not given the opportunity to fill any vacancy within the teacher team using incentive funds. Instead, principals filled vacancies using existing recruitment and hiring methods (19% were new hires, 22% were transfers from other schools, 30% moved from another position within the school, and the remaining vacancies either were not filled or had an unknown status). Students in the comparison condition could have been taught by a new teacher that filled a vacancy (focal teacher) or an existing teacher (nonfocal teacher).[5]

Outcomes and measurement

To measure mathematics and reading achievement, state-specific assessments were converted to *z*-scores to measure student achievement relative to the average performance within a particular state. Outcomes were measured at the end of 1 year following random assignment for the full randomized sample, and at the end of 2 years following random assignment for a subset of students from the first of two cohorts that entered the sample. For a more detailed description of these outcome measures, see Appendix B.

Support for implementation

Transferring teachers were given a brief orientation.

Reason for review

This study was identified for review by the WWC because it received significant media attention.

Appendix B: Outcome measures for each domain

Mathematics achievement	
Mathematics assessment	State-specific mathematics assessments were converted to *z*-scores to measure student achievement relative to the average performance within a particular state. The *z*-scores were calculated by taking a student's scaled score, subtracting it from the statewide mean scaled score for all students in that year and grade, and dividing the result by the statewide standard deviation of scaled scores.
Reading achievement	
Reading assessment	State-specific reading assessments were converted to *z*-scores to measure student achievement relative to the average performance within a particular state. The *z*-scores were calculated by taking a student's scaled score, subtracting it from the statewide mean scaled score for all students in that year and grade, and dividing the result by the statewide standard deviation of scaled scores.

Table Notes: Based on information provided by the authors, the WWC determined that results measuring effects on teacher retention rates do not meet WWC group design standards due to lack of equivalence of the analytic sample. The study considered the fall following random assignment as the baseline measure of "full retention." However, random assignment of teacher teams to condition occurred between the prior April through August. The WWC determined that demonstration of equivalence was therefore necessary based on evidence that the composition of teacher teams changed between random assignment and the beginning of the school year (p. 42 and Appendix A.3 of the report), and the study did not demonstrate the equivalence of the teacher teams for this analysis. The WWC also considered a number of intermediate outcomes to be ineligible to review, including: the method of assignment of teachers to students and grades, teacher mentoring and leadership roles, teacher attitudes, and principal reports on school climate.

Appendix C: Study findings for each domain

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Elementary mathematics achievement								
Mathematics assessment, program Year 1 (Cohorts 1 and 2)	Elementary students	97 teams/ 6,253 students	–0.27 (0.96)	–0.32 (1.02)	0.05	0.05	+2	.16
Mathematics assessment, program Year 2 (Cohort 1 only)	Elementary students	90 teams/ 6,139 students	–0.17 (1.02)	–0.27 (0.98)	0.10	0.10	+4	.01
Domain average for elementary mathematics achievement						**0.08**	**+3**	**Statistically significant**
Middle school mathematics achievement								
Mathematics assessment, program Year 1 (Cohorts 1 and 2)	Middle school students	30 teams/ 8,038 students	–0.56 (0.87)	–0.54 (0.88)	–0.02	–0.02	–1	.72
Mathematics assessment, program Year 2 (Cohort 1 only)	Middle school students	13 teams/ 2,355 students	–0.36 (0.89)	–0.35 (0.95)	–0.01	–0.01	0	.80
Domain average for middle school mathematics achievement						**–0.02**	**–1**	**Not statistically significant**
Elementary reading achievement								
Reading assessment, program Year 1 (Cohorts 1 and 2)	Elementary students	97 teams/ 6,200 students	–0.37 (0.95)	–0.41 (0.98)	0.04	0.04	+2	.27
Reading assessment, program Year 2 (Cohort 1 only)	Elementary students	90 teams/ 6,103 students	–0.28 (1.00)	–0.38 (0.98)	0.10	0.10	+4	< .01
Domain average for elementary reading achievement						**0.07**	**+3**	**Statistically significant**
Middle school reading achievement								
Reading assessment, program Year 1 (Cohorts 1 and 2)	Middle school students	31 teams/ 7,063 students	–0.53 (0.92)	–0.57 (0.88)	0.04	0.04	+2	.19
Reading assessment, program Year 2 (Cohort 1 only)	Middle school students	20 teams/ 3,128 students	–0.45 (0.99)	–0.43 (0.95)	–0.02	–0.02	–1	.29
Domain average for middle school reading achievement						**0.01**	**0**	**Not statistically significant**

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 individual 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 WWC-computed average effect size is a simple average

rounded to two decimal places; the average improvement index is calculated from the average effect size. The statistical significance of the study's domain average was determined by the WWC. Some statistics may not sum as expected due to rounding.

Study Notes: No corrections for clustering were needed because the study authors adjusted standard errors in their analyses based on clustering at the teacher team level (i.e., the level of random assignment). A correction for multiple comparisons was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. The WWC presents results from the author-reported analyses that exclude students who did not have a pretest score. These results differ slightly from the results the authors reported in their benchmark imputed analysis. For the elementary school analyses, the study is characterized as having a statistically significant positive effect in both the mathematics achievement domain and the reading achievement domain because the effect for at least one measure within each domain is positive and statistically significant and no effects are negative and statistically significant, accounting for multiple comparisons within each domain. For the middle school analyses, the study is characterized as having an indeterminate effect in both the mathematics achievement domain and the reading achievement domain because, within each domain, the mean effect reported is neither statistically significant nor substantively important. For more information, please refer to the WWC Standards and Procedures Handbook (version 3.0), pp. 25–26.

Appendix D: Supplemental findings by domain

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Elementary mathematics achievement								
Mathematics assessment, program Year 1 for students of focal teachers (Cohorts 1 and 2)	Elementary students	96 teams/ 2,834 students	–0.24 (0.94)	–0.42 (0.98)	0.18	0.19	+7	< .01
Mathematics assessment, program Year 2 for students of focal teachers (Cohort 1 only)	Elementary students	90 teams/ 2,856 students	–0.10 (0.99)	–0.31 (0.93)	0.21	0.22	+9	< .01
Mathematics assessment, program Year 1 for students of nonfocal teachers (Cohorts 1 and 2)	Elementary students	87 teams/ 4,961 students	–0.30 (0.97)	–0.27 (1.07)	–0.03	–0.03	–1	.51
Mathematics assessment, program Year 2 for students of nonfocal teachers (Cohort 1 only)	Elementary students	82 teams/ 4,957 students	–0.17 (1.03)	–0.24 (1.01)	0.07	0.07	+3	.11
Middle school mathematics achievement								
Mathematics assessment, program Year 1 for students of focal teachers (Cohorts 1 and 2)	Middle school students	30 teams/ 2,533 students	–0.53 (0.83)	–0.58 (0.85)	0.05	0.06	+2	.54
Mathematics assessment, program Year 2 for students of focal teachers (Cohort 1 only)	Middle school students	13 teams/ 1,403 students	–0.40 (0.87)	–0.43 (0.89)	0.03	0.03	+1	.46
Mathematics assessment, program Year 1 for students of nonfocal teachers (Cohorts 1 and 2)	Middle school students	30 teams/ 7,754 students	–0.65 (0.86)	–0.60 (0.84)	–0.05	–0.06	–2	.38
Mathematics assessment, program Year 2 for students of nonfocal teachers (Cohort 1 only)	Middle school students	13 teams/ 1,621 students	–0.38 (0.87)	–0.34 (0.96)	–0.04	–0.04	–2	.59

Elementary reading achievement								
Reading assessment, program Year 1 for students of focal teachers (Cohorts 1 and 2)	Elementary students	96 teams/ 2,864 students	–0.40 (0.92)	–0.54 (0.97)	0.14	0.15	+6	< .01
Reading assessment, program Year 2 for students of focal teachers (Cohort 1 only)	Elementary students	88 teams/ 2,742 students	–0.26 (0.95)	–0.47 (0.91)	0.21	0.23	+9	< .01
Reading assessment, program Year 1 for students of nonfocal teachers (Cohorts 1 and 2)	Elementary students	91 teams/ 5,021 students	–0.40 (0.99)	–0.40 (0.99)	0.00	0.00	0	.93
Reading assessment, program Year 2 for students of nonfocal teachers (Cohort 1 only)	Elementary students	88 teams/ 5,509 students	–0.30 (1.04)	–0.34 (1.00)	0.04	0.04	+2	.24
Middle school reading achievement								
Reading assessment, program Year 1 for students of focal teachers (Cohorts 1 and 2	Middle school students	31 teams/ 2,916 students	–0.54 (0.90)	–0.57 (0.89)	0.03	0.03	+1	.50
Reading assessment, program Year 2 for students of focal teachers (Cohort 1 only)	Middle school students	20 teams/ 1,841 students	–0.45 (0.90)	–0.39 (0.91)	–0.06	–0.07	–3	.05
Reading assessment, program Year 1 for students of nonfocal teachers (Cohorts 1 and 2)	Middle school students	31 teams/ 6,491 students	–0.57 (0.92)	–0.62 (0.86)	0.05	0.06	+2	.12
Reading assessment, program Year 2 for students of nonfocal teachers (Cohort 1 only)	Middle school students	20 teams/ 2,619 students	–0.46 (1.03)	–0.42 (0.94)	–0.04	–0.04	–2	.12

Table Notes: The supplemental findings presented in this table are additional findings that do not factor into the determination of the evidence 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 individual 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.

Study Notes: The study presented impacts separately for focal teachers and nonfocal teachers within each teacher team. Teachers in both the *TTI* and comparison conditions who filled vacant positions were designated as the focal teachers, and all other teachers in the team were designated as nonfocal. The WWC did not factor results from the focal and nonfocal teacher results into the rating of effectiveness due to ambiguities in identifying which teachers were focal and nonfocal. See Endnote 4 on p. 10 for a more detailed description of why the WWC rated these analyses as *meets WWC group design standards with reservations*. No corrections for clustering or multiple comparisons and no difference-in-differences adjustments were needed. The *p*-values presented here were reported in the original study. WWC-calculated effect sizes differ slightly from the standardized mean differences reported by the authors. None of the differences have bearing on whether an impact would be deemed substantively important by the WWC.

Endnotes

[1] Single study reviews examine evidence published in a study (supplemented, if necessary, by information obtained directly from the authors) to assess whether the study design meets WWC group design standards with or without reservations. The review reports the WWC's assessment of whether the study meets WWC group design standards and summarizes the study findings following WWC conventions for reporting evidence on effectiveness. This study was reviewed using the single study review protocol (version 2.0). A quick review of this study was released in February 2014, and this report is the follow-up review that replaces that initial assessment. The WWC rating applies only to the study outcomes that were eligible for review under this topic area. The reported analyses in this SSR are only for those eligible outcomes that either met WWC group design standards without reservations or met WWC group design standards with reservations, and do not necessarily apply to all results presented in the study.

[2] Absence of conflict of interest: This study was conducted by staff from Mathematica Policy Research. Because Mathematica is one of the contractors that administers the WWC, this study was reviewed by staff from subcontractor organizations.

[3] There was one eligible outcome, teacher retention rate, included in the study that is not described in this WWC report because analyses for this outcome do not meet WWC group design standards. See the table notes in Appendix B for more information. The WWC also considered a number of intermediate outcomes to be ineligible to review, including: method of assignment of teachers to students and grades, teacher mentoring and leadership roles, teacher attitudes, and principal reports on school climate.

[4] In addition to presenting results for the full sample, the study also presented impacts separately for the "focal teachers" (i.e. those teachers within a teacher team who were filling the vacant slot) and "nonfocal" teachers (the other teachers in the team). All reading and mathematics analyses measuring impacts for the subset of focal and nonfocal teachers within teams *meet WWC group design standards with reservations*. These results are presented in Appendix D and are not included in the rating of program effectiveness. The study authors acknowledged that there was ambiguity in designating teachers as focal or nonfocal, particularly within the comparison group. This uncertainty occurred in 23% of comparison teams and 5% of *TTI* teams (see p. 42 of report). The WWC determined that there was reason to be concerned that the samples in the analysis were not representative of the team-level randomization process and, therefore, determined that these impact results are not eligible for a rating of *meets WWC group design standards without reservations* and should not be included in the rating of program effectiveness. However, these comparisons *meet WWC group design standards with reservations* because the study established baseline equivalence of the *TTI* and comparison samples. Since these samples are effectively a subgroup of the full sample analyses, they are presented in Appendix D.

[5] Impacts for focal and nonfocal teachers are presented in Appendix D. See Endnote 4 above for an explanation of the study rating of *meets WWC group design standards with reservations* for these subgroup results.

Recommended Citation

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2015, March). *WWC review of the report: Transfer incentives for high-performing teachers: Final results from a multisite randomized experiment.* Retrieved from http://whatworks.ed.gov

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 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.

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.

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.

Single-case design (SCD) 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.

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 are 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.

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