The Implementation and Effects of a Multi-year Continuous Improvement Research Project in Rural Secondary Schools: Montana Partnership Comparison Study 2

Grant Number R305H150003

August 2020

Prepared for:
James Benson
Institute of Education Sciences
550 12th Street SW
Washington, DC 20202
james.benson@ed.gov

Prepared by:
SRI International
Todd Grindal
Stephanie Nunn
Xin Wei
Jared Boyce
Kirby Chow
SRI Project Number: P23253

SRI Education™
A DIVISION OF SRI INTERNATIONAL

Background

This report presents information on the relationship between participation in the Montana Partnership project (MTP) and the literacy skills of students and instructional practices of teachers in participating schools. Supported by a Continuous Improvement Research in Education grant from the Institute of Education Sciences, MTP was a 4-year collaboration between SRI Education (SRI) and the Montana Office of Public Instruction (OPI) focused on improving the literacy skills of middle and high school students who struggle with literacy/reading.

Two middle schools and two high schools participated in the project. The Montana OPI nominated these schools for participation because they had a large percentage of students who did not meet state literacy standards. Within each participating school, a team of 8–12 school administrators, instructional coaches, reading intervention teachers, and general education teachers collaborated to implement a set of quick turnaround interventions using a Plan Do Study Act (PDSA) process. These PDSA teams worked with SRI and OPI to design and implement two 6-week PDSA cycles in each school year.

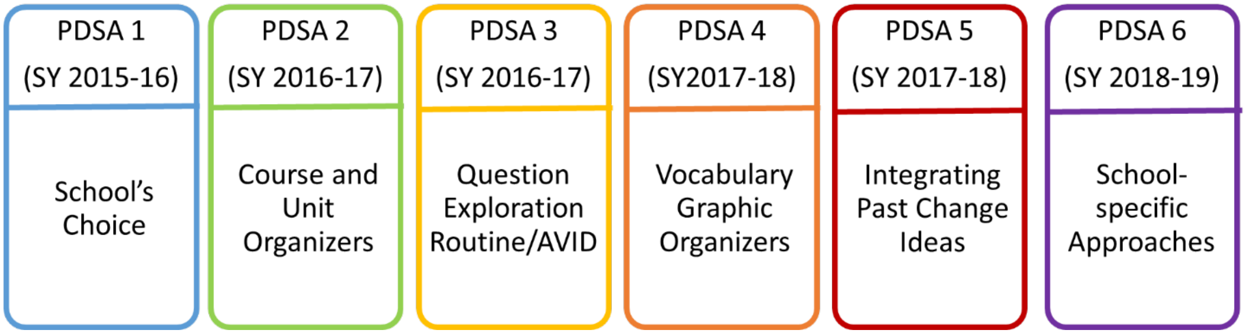
During each PDSA cycle, team members implemented small alterations in their school practices or processes, referred to as a “change idea,” which is designed to advance student literacy skills and focused on students who struggled with reading. PDSA team members were encouraged to select change ideas that would support struggling readers in particular, with the idea that these supports would be help all learners access core classroom instruction and content. Teams met weekly to review data on relevant teacher practices and make changes to implementation. At the end of the cycle, the teams review the data on student outcomes and develop a plan for the next PDSA cycle. School teams engaged in six PDSA cycles across the 4-year project.

- PDSA cycle 1 (2015–16 school year): Each school was encouraged to identify a specific change idea using the school’s instructional framework. PDSA strategies included implementing formative assessments, such as exit tickets, in each class period; focusing a vocabulary strategy on prefixes and suffixes; and using ability groups during in-class writing activities.
- PDSA cycles 2 and 3 (2016–17 school year) focused on the Course and Unit Organizer Routines [The Strategic Instruction Model (SIM) strategies] (PDSA 2), the Question Exploration Routine (SIM strategy), and a similar Advancement Via Individual Determination (AVID) strategy (PDSA 3).
- PSDA cycle 4: (2017–18 school year) school teams implemented two additional PDSA cycles. PDSA 4 focused on use of school-specific instructional strategies for supporting vocabulary skill development [LINCS (SIM), Key ideas, Information, Memory clue (KIM), 4-Square].
- PDSA 5 (2017–18 school year) then focused on the simultaneous integration of these vocabulary strategies with the strategies from the 201617 school year. In the fourth year of

the project, school teams expressed a desire for more autonomy and choice over the strategies, and additional time for implementation and testing.

- PDSA 6, the final PDSA cycle, took place over 12 weeks during the 2018–19 school year. During PDSA 6, the school teams chose strategies that include: one-sentence summaries, growth mindsets, checks for understanding, and weekly student-teacher conferences. The project's PDSA cycles and their respective change ideas are shown below (Figure 1).

Figure 1 PDSA Cycle Topics in Montana, School Years 2015-16 through 2018-19



PDSA 1 (SY 2015-16)	PDSA 2 (SY 2016-17)	PDSA 3 (SY 2016-17)	PDSA 4 (SY2017-18)	PDSA 5 (SY 2017-18)	PDSA 6 (SY 2018-19)
School's Choice	Course and Unit Organizers	Question Exploration Routine/AVID	Vocabulary Graphic Organizers	Integrating Past Change Ideas	School-specific Approaches

In this report, we use descriptive data along with quasi-experimental analyses to compare student literacy skills and teacher practices in the MTP treatment schools with those of students and teachers in a set of similar non-participating comparison schools. Whenever possible, our analyses focus in particular on students who struggle with reading. In addition, these analyses describe the treatment schools' implementation of the intervention activities from the Plan Do Study Act (PDSA) cycle change ideas.

We examine the following research questions

Student-focused research questions:

(1) What differences in the fall 2017 to spring 2018 gains, as measured by ISIP reading scores, are observed between students reading at or above grade level (Tier 1 students) who received PDSA cycles in the treatment middle schools and similar students in matched comparison schools?

(2) What differences in the fall 2017 to spring 2018 gains, as measured by ISIP reading scores, are observed between students reading below or significantly below grade level (Tiers 2 and 3 students) who received PDSA cycles in treatment middle schools and similar students in matched comparison schools?

(3) What differences measured by SBAC ELA are observed between students who received PDSA cycles in treatment middle schools and their peers in matched comparison middle schools after each PDSA cycle across 3 years of implementation?

(4) What differences measured by ACT Reading and English assessments are observed between students who received PDSA cycles in treatment high schools and similar peers in matched comparison high schools across 3 years of implementation?

School-leader focused research questions:

(1) What differences are reported on the school survey by school leaders between treatment schools and comparison schools in the school's allocation of time for students in Tiers 1, 2, and 3 at the middle school level across 3 years of implementation?

(2) What differences are reported on the school survey by school leaders between treatment schools and comparison schools in the school's allocation of time for students in Tiers 1, 2, and 3 at the high school level across 3 years of implementation?

Teacher-focused research questions:

(1) What differences do treatment teachers and comparison teachers report on instructional practices at the middle school level across 3 years of implementation?

(2) What differences do treatment teachers and comparison teachers report on instructional practices at the high school level across 3 years of implementation?

(3) How often do middle and high school treatment teachers report student performance (exit tickets) during a PDSA cycle across 3 years of implementation?

(4) What is the amount of time that middle and high school treatment teachers report providing literacy intervention activities during each PDSA cycle? How does the amount differ across teachers and schools?

Measures and Analytic Methods

Data

We used three types of data to answer the research questions. Student outcome data from the treatment schools and comparison schools were used to answer research questions 1–4 regarding differences in student literacy/reading achievement. Teacher and school leader self-report survey data were used to address research questions 5–6 regarding instructional time allocated to literacy/reading instruction. Data from teacher and school leader surveys were used to address research questions 7–10 regarding differences in literacy/reading teacher instruction, intervention, and assessment.

Student outcome data

ISIP Advanced Reading. During the study period, Montana secondary schools used ISIP Advanced Reading to assess students' overall reading ability and their fluency, decoding, comprehension, and vocabulary skills. ISIP Advanced Reading is a computer adaptive test which is administered individually and aligned with the Common Core State Standards (CCSS). In addition, ISIP Advanced Reading has established concurrent validity with the Gray Oral Reading Test-4 (GORT-4), Woodcock-Johnson-3 (WJ-III), Wechsler Individual Achievement Test-II (WIAT-II) and the Peabody Picture Vocabulary Test-IV (PPVT-IV) (Mathes, 2016). ISIP Advanced Reading provides an ability score, as well as percentile ranks and Tier placement information, for overall reading and for each subscale. Because the assessment is computer adaptive, the software uses marginal reliability to establish and maintain internal consistency. ISIP Advanced Reading has a stopping criterion based on minimizing the standard error of the ability score. As such, the lower limit of the marginal reliability of the data for any testing instance will always be approximately 0.90 (Mathes, 2016).

The intervention and comparison middle and high schools had been using this assessment to monitor student progress and identify students in need of additional supports since 2011. There are two drawbacks to using ISIP Advanced Reading as an outcome for this study. First, ISIP Advanced Reading is marketed for use for students in fourth through eighth grades and the norms and standards of the program were developed using students from this age range. Secondly, the norming data used to develop the ISIP Advanced Reading norms are based on a majority White, Latino/a, and African American sample, with less than 1% of the sample identifying as American Indian (Mathes, 2016). In six of the intervention and comparison middle and high schools a majority of the enrolled students identify as American Indian, and thus there are concerns about the validity of this assessment for this group of students. However, this is an issue with the majority of available standardized assessments, as most have limited data from American Indian populations. For instance, the ACT and WJ-III were both normed using a sample containing less than 1% American Indian high school students (ACT, 2019; Mather & Woodcock, 2001). The schools' familiarity with and use of ISIP Advanced Reading, the frequency with which the assessment can be administered, and the continued involvement of ISIP with the state to improve the test with both high school and American Indian students contributed to our decision to select this test as a student-level reading outcome.

MT schools were expected to administer the assessment to all enrolled students in September, December, and May of each school year. In addition, students with reading scores in the Tier 2 (performing slightly below grade level) and Tier 3 (performing far below grade level) range were expected to take the assessment monthly through the school year. We were able to obtain ISIP Advanced Reading data from both treatment and comparison schools from the 2016–17 and 2017–18 school years. Following the 2017–18 school year, comparison schools transitioned to new progress monitoring assessments that better fit their needs and thus could no longer provide ISIP data for this study.

SBAC ELA. Montana's statewide assessment of student learning is the Smarter Balanced Assessment Consortium (SBAC). The SBAC is a Common Core Standards Initiative-aligned, computer-adaptive test that measures students' college and career readiness in English language arts (ELA) and math. In addition to raw and scaled scores, performance levels are reported in 4 levels for accountability purposes. Montana calculated scaled scores for students' performance on the SBAC reading and the SBAC ELA and assigned students to a corresponding reading proficiency level from 1 to 4, with 1 indicating Novice skills, 2 indicating Nearing Proficiency, 3 indicating Proficiency, and 4 indicating Advanced Proficiency. In creating the SBAC ELA, the consortium used Standards for Educational and Psychological Testing (AERA et al., 2014) to assess the validity of the measure along a variety of indicators (SBAC Technical Report, 2016). Ongoing plans to measure the validity and reliability of the measures have been established (SBAC Technical Report, 2016) with reliability alpha ranges from 0.69 to 0.81 on ELA (SBAC Technical Report, 2016). Montana students in the third through eighth grades take the statewide assessment each spring

ACT Reading and English. Beginning in the 2011–12 school year, Montana began requiring all high school juniors to register for and take the ACT college readiness assessment. The ACT test includes two subtests which are of interest for this study: Reading and English. These two subtests consist of a series of multiple-choice questions which are scored for accuracy. This raw score is then converted to a scale score ranging from 1 to 36.

The ACT English Test provides students 45 minutes to answer 75 multiple-choice questions about mechanics, usage, and rhetoric. The test is divided into five passages, each with approximately 15 questions. The test is designed to measure students' understanding of the conventions of English, including punctuation, grammar, and sentence structure, as well as their rhetorical skills, including text organization and style. The ACT English Test is scored on a scale from 1 to 36, with a score of 18 (representing approximately the 40th percentile) indicating college and career readiness or proficiency.

The ACT Reading section consists of four, approximately 1000-word passages (one each of Social Studies, Natural Sciences, Humanities, and Literary Narrative or Prose Fiction) written at a college reading level. For each passage, students are asked to read the text and then respond to 10 multiple-choice questions. These questions ask students to recognize and identify the theme of the passage, comprehend and recall specific facts about the passage, understand the structure of the passage, and make inferences about each passage's theme. Students are given 35 minutes to complete this section of the test. The ACT Reading Section is scored on a scale from 1 to 36, with a score of 22 (representing approximately the 61st percentile) indicating college and career readiness or proficiency.

In terms of reliability, the score scales for the ACT test were developed to have approximately constant standard errors of measurement for all true scale scores. Assuming a normal distribution of measurement error, about two-thirds of students who take the ACT test can be expected to be mis-measured by less than 1 standard error of measurement. Additional detail about the reliability of the ACT test can be found in the assessment's technical manual (ACT, 2019). The content validity of the ACT test has been repeatedly demonstrated through surveys of high school curricula and standards (ACT, 2019). In addition, the ACT test has also showed construct validity through correlational studies comparing students' high school grades and ACT scores, as well as first-year college course grades and ACT scores (ACT, 2019). Finally, the variability of ACT test scores is not associated with race/ethnicity or gender (< 2% of total variability in test scores; Noble et al., 1999), an important consideration for the current study's emphasis on the performance of American Indian students.

We received ACT data from all treatment and comparison high schools from the 2016–17, 2017–18, and 2018–19 school years.

School leader and teacher self-report surveys

The SRI team administered online surveys to gather information from teachers and school leaders in treatment and comparison schools. In these surveys, teachers were asked to report their expectations and practices for addressing the literacy needs of Tier 2 and 3 students. The surveys targeted three different types of staff in each school: content area teachers, interventionist teachers, and school leaders. The surveys ranged in length from 13 to 29 questions depending on the respondent's role and used a Likert scale (1 to 7) response format.

Selection of Schools for the Comparison Group

Comparison schools were selected from the 11 other Montana middle and high schools that had participated in the Montana Striving Readers Project, a previous literacy focused school improvement initiative. We applied nearest neighbor propensity score matching to select four comparison schools similar to the four treatment schools based on literacy and demographic measures collected during the 2014–15 school year.[1] The matching procedure was done separately for middle and high schools. The nearest neighbor propensity score approach identified the comparison schools that had the closest propensity score to each treatment school (Table 1).[2]

[1] Details on the school matching procedure are in Comparison Report 1 submitted to OPI on September 30, 2016.

[2] A propensity score is the predicted probability of participating in an intervention based on a set of potentially confounding covariates (school-level student literacy achievement and school characteristics) using logistic regression. There was a pool of six nonintervention middle schools and five nonintervention high schools from which to select comparison schools. To start, we posited a logistic model to estimate what types of schools were likely to be the intervention schools using school-level variables. Because the sample size is very small ($n = 8$ for middle schools and $n = 7$ for high schools), we included only two predictors in our logistics models to avoid overfitting of the logistic model. We calculated a propensity score (logit) of being an intervention school based on literacy score in 2014─15 and attendance rate in 2014─15. We selected comparison schools that were closest to each intervention school on the propensity score using nearest neighbor matching.

Table 1 Treatment and Comparison Schools

Treatment Schools	Comparison Schools
Hardin Middle School	Libby Middle School
Browning Middle School	Charlo Middle School
	East Middle School
Browning High School	Hardin High School
Anaconda High School	Wolf Point High School
	Libby High School

Propensity score methods are set of quasi-experimental approaches were developed to approximate findings obtained from randomized control trials (Becker & Ichino, 2002). They have been increasingly used in observational studies with cohort designs to reduce selection bias in estimating treatment or intervention effects when randomized controlled trials are not feasible or ethical (Rosenbaum & Rubin, 1983, 1984, 1985).

We also propensity scores to test the effect of participation in the PDSA cycles on student literacy outcomes. The propensity score is the predicted probability of participating in a treatment based on a set of potentially confounding covariates (i.e., student demographic and disability characteristics, baseline score) using logistic regression. Propensity scoring attempts to equalize the mean values of potentially confounding observed covariates in the treatment and comparison groups, assuring that differences in outcomes between the treatment and covariate effect are not the result of differences in mean values of those covariates.

We estimated the average treatment effect on the treated (ATT) (Curtis, Hammill, Eisenstein, Kramer, and Anstrom (2007); Hirano, Imbens, and Ridder (2003); and Rosenbaum and Rubin (1983). Specifically, the weight for treated students was 1.0 and the weight for nonparticipating students was equal to (pi/1-pi), where pi is the propensity score for the i-th comparison student (Harder, Stuart, & Anthony, 2010; Hirano et al., 2003).

Analysis

For the analysis of student-level research questions (RQ1– 4), we used propensity score weighted Hierarchical Linear Modeling techniques (HLM, Raudenbush & Bryk, 2002) to estimate the impact of intervention on student literacy achievement. HLM takes into account that (a) repeated measures from the same students are correlated, and (b) treatment and control students are nested within teachers. In the example below, outcomes are May iSIP Advanced Reading scores. Covariates in the model include pretest scores (September iSIP Advanced Reading scores) in addition to student gender, special education status, and race variables that may reduce residual variability.

The 2-level propensity score weighted HLM model for treatment effects is as follows:

$Yij=\beta0+\beta1Intervention+\beta2(COV-COV..)+\mu0j+eij$

Yij is May iSIP reading scores of student *i* in teacher *j*. *Intervention* indicates initial assignment with 1 for intervention and 0 for control. The coefficient β_1 associated with *intervention* in the above HLM model indicates the average treatment effect in promoting improved student iSIP Advanced Reading scores. β_2 are coefficients associated with each covariate including student baseline demographic characteristics (e.g., gender and race) and September iSIP Advanced Reading scores in the fall.

eij

is student random effect, and

μ0j

is teacher random effect which will facilitate interpreting results as applying to a potentially wider universe of teachers than the particular teachers in this study. We will run this analysis using all students as well as a subsample of Tiers 2 and 3 students.

Research Questions 1 and 2: Sample and Results

Data and Data Cleaning for ISIP Outcomes Data

Several issues limit our ability to address the research questions, including inconsistent data collection and missing data. The results reported here use complete case analyses of only students with data for all the required variables. These issues substantially reduce the statistical power of this study. Given the size of the analytic sample in the treatment schools (798 students in four schools), students in treatment schools would need to demonstrate fall-to-spring literacy skill gains that are more than one standard deviation greater than those of students in the comparison schools for the results to be statistically significant at $p < .05$. We therefore considered the magnitude of non-significant effects when interpreting these results.

Table 2 presents the sample sizes for the maximal sample of students with any amount of data and the analytic sample for the complete case analysis in 2016–17 (Project Year 2) and Table 3 presents these data for 2017–2018 (Project Year 3).

Table 2 Sample Sizes for ISIP Advanced Reading Analyses, 2016–17 (Project Year 2)

Variable	Sample (Grades 7 and 8)	Sample (Grades 7 and 8 with no missing ISIP or demographic data)
MTP treatment schools	2282	2076
Browning Middle School	1508	1442
Hardin Middle School	774	634
Comparison schools	1443	901
Libby Middle School	195	166
Wolf Point High School	1248	735
Total	3725	2977

Table 3 Sample Sizes for ISIP Advanced Reading Analyses, 2017–18

Variable	Sample (Grades 7 and 8)	Sample (Grades 7 and 8 with no missing ISIP or demographic data)
MTP treatment schools	576	440
Browning Middle School	283	208
Hardin Middle School	293	232
Comparison schools	763	674
Charlo 7–8	43	28
East Middle School	730	646
Total	1339	1114

Table 4 presents descriptive statistics for students in the treatment and comparison schools for the complete case analysis sample in 2016–17. The samples are largely similar in terms of the percentage of students qualifying for special education services (14% vs. 10%). Most of the students in treatment schools identify as American Indian (93%), whereas a smaller majority of students in the comparison schools identify as American Indian (56%). Notably, students in treatment schools on average scored 146.8 points lower (-0.75 effect size) on the baseline ISIP assessment than did comparison school students. After propensity score weighting, the two groups were similar on all demographic characteristics and baseline ISIP scores.

Table 4 Complete Case Analytic Samples for Treatment and Comparison Schools on ISIP Advanced Reading, 2016–17

Variable	Treatment school (*N* = 1882)		Comparison Schools (*N* = 810)		Cohen's d size difference before propensity score weighting	Weighted Comparison schools (*N* = 810)		Cohen's d size difference after propensity score weighting
	Mean	*SD*	Mean	*SD*		Mean	*SD*	
Grade 8	0.48	0.50	0.44	0.50	0.08	0.41	0.73	0.14
Gender (1 = male)	0.54	0.50	0.53	0.50	0.02	0.52	0.75	0.04
Race/Ethnicity								
American Indian	0.93	0.26	0.56	0.48	1.08	0.90	0.39	0.09
Other	0.02	0.15	0.15	0.35	-0.57	0.03	0.23	-0.04
White (reference)	0.05	0.22	0.28	0.41	-0.79	0.07	0.32	-0.07
Special education (1 = Yes)	0.14	0.34	0.10	0.31	0.12	0.09*	0.42	0.15
Baseline tier	2.40	0.80	1.81	0.88	0.72	2.37	1.21	0.04
Baseline score	1993.88	192.08	2140.67	206.87	-0.75	2002.90	295.50	-0.05
May score	2053.42	235.34	2207.45	220.53	-0.67	2083.2	291.7	-0.13

Note. Four schools are included in this analysis. They are Libby Middle school, Wolf Point high school, Browning Middle school, Hardin Middle School. Limited English Proficiency status and Economically Disadvantaged status were not included in the analysis because of high levels of missing data.

Table 5 presents descriptive statistics for students in the treatment and comparison schools for the complete case analysis sample in 2017–18. There are several notable differences between the treatment and comparison schools. Most of the students in treatment schools identify as American Indian (90%), whereas those in the comparison schools identify as White (75%). A higher percentage of treatment school students are of limited English proficiency than comparison students (16% vs. 4%) and are economically disadvantaged (100% vs. 60%). Further, students in treatment schools on average scored 125.2 points lower (-0.68 effect size) on the baseline ISIP assessment than did comparison school students. After propensity score weighting, the two groups were similar on all demographic characteristics and baseline ISIP scores.

Table 5 Complete Case Analytic Samples for Treatment and Comparison Schools on ISIP Advanced Reading, 2017–18

Variable	Treatment school (*N* = 1882)		Comparison Schools (*N* = 674)		Cohen's d size difference before propensity score weighting	Weighted Comparison schools (*N* = 674)		Cohen's d size difference after propensity score weighting
	Mean	***SD***	**Mean**	***SD***		***Mean***	***SD***	
Grade 8	0.51	0.50	0.49	0.50	0.04	0.49	0.40	0.04
Gender (1 = male)	0.49	0.50	0.50	0.50	-0.02	0.48	0.40	0.02
Race/Ethnicity								
American Indian	0.90	0.30	0.18	0.39	2.18	0.90	0.24	0
Other	0.03	0.16	0.07	0.26	-0.20	0.02	0.09	0.05
White (reference)	0.08	0.26	0.75	0.43	-2.09	0.09	0.23	-0.03
Special education (1 = Yes)	0.09	0.29	0.09	0.29	0	0.11	0.25	-0.07
Limited English Proficiency (1 = Yes)	0.16	0.37	0.04	0.19	0.37	0.20	0.32	-0.12
Economically disadvantaged (1 = Yes)	1	0	0.60	0.49	1.49	1	0.0008	0
Baseline tier	2.19	0.84	1.67	0.81	0.63	2.13	0.67	0.07
Baseline score	2044.73	180.37	2169.89	191.77	-0.68	2053.86	141.58	-0.05
May score	2139.41	211.68	2230.92	182.24	-0.45	2136.83	134.49	0.01

Findings Research Questions 1 and 2

Propensity score weighted regression analyses suggest that students in the comparison schools, on average, experienced greater growth in their reading skills from September 2016 to May 2017 as compared to students in schools that participated in the MTP. These differences are not statistically significant, meaning that we cannot determine whether what we observe in these analyses is the result of a true difference between the groups or the result of chance.

Table 6 Complete Case Analysis, Propensity Score Weighted Treatment Main Effect, Grades 7–8, 2016-17

Variable	All Students *N* = 1882 Treatment vs. 810 Comparison	Tier 1 *N* = 371 Treatment vs. 406 Comparison	Tiers 2 & 3 *N* = 1511 Treatment vs. 404 Comparison
Intercept	364.94* (38.29)	458.13* (92.47)	423.18* (62.24)
Treatment (1 = MTP)	**-51.59 (35.38)**	**-3.52 (18.42)**	**-76.36 (58.42)**
Grade 8	29.69*** (5.12)	6.69 (8.48)	40.79*** (6.61)
September baseline test	0.89 *** (0.01)	0.83*** (0.04)	0.85*** (0.03)
Gender (1 = Male)	-23.57*** (4.86)	0.48 (8.23)	-30.40*** (5.86)
Race/Ethnicity			
American Indian	-26.26* (13.31)	-28.49* (13.32)	-0.23 (22.71)
Other	-47.86* (19.65)	-25.26 (27.86)	-29.84 (28.64)
Special education (1 = Yes)	-13.13 (8.41)	-100.79** (34.34)	14.86 (9.45)

* $p < .05$, ** $p < .01$

Data from 2017–18 do not provide strong evidence of a main effect of the intervention (Table 7). For Tiers 2 and 3, students in Grades 7 and 8 in treatment schools demonstrated spring literacy skills that were on average higher than those of students in comparison schools when controlling for demographic characteristics and fall test scores, although these differences did not reach statistical significance. If we translate these data into effect sizes these differences were -0.12 for all students, -0.12 for Tier 1 students, and 0.03 for Tier 2 and Tier 3 students. Because of sample size limitations, however, we were unable to determine with any reasonable degree of certainty that these observed differences represent a true association between MTP participation and student literacy skills.

Table 7 Complete Case Analysis, Propensity Score Weighted Treatment Main Effect, Grades 7–8, 2017–18

Variable	All Students *N* = 440 Treatment vs. 674 Comparison	Tier 1 *N* = 123 Treatment vs. 356 Comparison	Tiers 2 & 3 *N* = 317 Treatment vs. 190 Comparison
Intercept	307.84 (3902.34)	392.55 (3461.22)	348.38 (8990.12)
Treatment (1 = MTP)	**-24.90 (42.56)**	**-25.88 (51.22)**	**7.31 (16.30)**
Grade 8	7.98 (5.98)	-22.56* (9.53)	23.54** (8.23)
September baseline test	0.90**	0.88**	0.86**

Variable	All Students *N* = 440 Treatment vs. 674 Comparison	Tier 1 *N* = 123 Treatment vs. 356 Comparison	Tiers 2 & 3 *N* = 317 Treatment vs. 190 Comparison
	(0.02)	(0.05)	(0.04)
Gender (1 = Male)	-6.17 (5.86)	-0.96 (8.52)	-5.52 (8.06)
Race/Ethnicity			
American Indian	11.01 (10.98)	15.55 (11.73)	6.64 (18.86)
Other	-7.58 (23.47)	4.13 (24.14)	-23.64 (41.17)
Special education (1 = Yes)	6.61 (10.70)	-8.61 (34.16)	1.61 (13.28)
Limited English Proficiency (1 = Yes)	-19.17** (8.64)	-26.78 (40.12)	-23.53* (10.30)
Economically disadvantaged (1 = Yes)	9.26 (3901.95)	0.65 (3459.14)	11.09 (8989.80)

* $p < .05$, ** $p < .01$

Research Question 3: Sample and Results

Data and Data Cleaning for SBAC Outcomes Data

OPI provided SRI with deidentified SBAC data for 2016–17, 2017–18, and 2018–19 school years. Because student ID was not provided, we were not able to link students over multiple years. 3 Therefore, the analyses conducted as part of this study treat each school year of data as independent samples combined into a larger dataset. This means that although many students are measured in Grade 7 and Grade 8, we are unable to account for that repeated measurement in the analyses. Table 8 provides sample sizes for partnership and comparison schools in spring 2017, 2018, and 2019.

Table 8 Sample Sizes for SBAC Analyses

Variable	Year	Sample (Grades 7 and 8)	Sample (Grades 7 & 8 with no missing SBAC or demographic data)
MTP Treatment schools	Total	1636	1574
Browning Middle School	2019	253	219
	2018	250	250
	2017	273	273
Hardin Middle School	2019	301	301
	2018	279	276
	2017	280	280
Comparison schools	Total	4766	4618
Charlo 7–8	2019	36	36

[3] The data use agreement between SRI and OPI requires that only deidentified student data be shared for the purposes of this study.

Variable	Year	Sample (Grades 7 and 8)	Sample (Grades 7 & 8 with no missing SBAC or demographic data)
	2018	28	28
	2017	40	40
East Middle School	2019	1497	1371
	2018	1324	1318
	2017	1314	1314
Libby Middle School	2019	199	183
	2018	168	168
	2017	160	160
Total		6402	6192

Findings Research Question 3

The primary differences between the samples of children in treatment and comparison schools is that the percentage of students who identify as American Indian or are economically disadvantaged is higher in the treatment group. However, after propensity score weighting, these two groups were equivalent on these two demographic variables. Note that baseline student level test scores were not available for the propensity score weighting and impact analysis.

Table 9 Descriptive Analysis of the SBAC Analytic Sample Before Propensity Score Weighting and After Propensity Score Weighting

Variable	Treatment school (*N* = 1574)		Comparison Schools before propensity score weighting (*N* = 4618)		Cohen's d size difference before propensity score weighting	Comparison Schools after propensity score weighting (*N* = 4618)		Cohen's d size difference after propensity score weighting
	Mean	*SD*	Mean	*SD*		Mean	*SD*	
School Year 2019	0.31	0.46	0.34	0.48	-0.06	0.30	0.27	0.02
School Year 2018	0.33	0.47	0.33	0.47	0.00	0.33	0.28	0.00
School Year 2017	0.35	0.13	0.33	0.47	0.07	0.37	0.29	-0.07
Grade 8	0.50	0.50	0.49	0.50	0.02	0.51	0.29	-0.02
Grade 7	0.50	0.50	0.51	0.50	-0.02	0.49	0.29	0.02
Gender (1 = male)	0.49	0.50	0.52	0.50	-0.06	0.49	0.29	0.00
Race/Ethnicity								
Hispanic	0.03	0.16	0.05	0.22	-0.11	0.03	0.10	0.00
African American	0	0	0.02	0.14	-0.26	0	0	0.00
American Indian	0.87	0.33	0.09	0.29	2.45	0.88	0.19	-0.03

Variable	**Treatment school (*N* = 1574)**		**Comparison Schools before propensity score weighting (*N* = 4618)**		**Cohen's d size difference before propensity score weighting**	**Comparison Schools after propensity score weighting (*N* = 4618)**		**Cohen's d size difference after propensity score weighting**
Asian	0	0	0.01	0.10	-0.18	0	0	0.00
Other	0.02	0.15	0.04	0.19	-0.12	0.02	0.09	0.00
Special education (1 = Yes)	0.08	0.28	0.11	0.31	-0.10	0.07	0.15	0.03
Economically disadvantaged (1 = Yes)	0.996	0.06	0.52	0.50	1.71	0.996	0.04	0.00
SBAC Reading Scale Score from 2016 to 2019	2454.99	120.38	2544.35	158.26	-0.67	2500.95	88.38	-0.35

The analyses do not provide evidence of an association between intervention and student performance on the SBAC (Table 10). Students in Grade 7, Grade 8, and both grades combined in treatment schools did not earn statistically significantly higher SBAC scores than those of students in the weighted comparison group after controlling for demographic characteristics. The lack of baseline SBAC data for analysis makes it difficult for us to determine with any reasonable degree of certainty whether the results of these analyses represent a true null association between MTP participation and student reading achievement.

Table 10 Propensity Score Weighted HLM Results for SBAC, 7th and 8th Grades Combined, 7th Grade, and 8th Grade

Variable	**7th and 8th Grades *N* = 1574 Treatment vs. 4618 Comparison**	**7th Grade *N* = 794 Treatment vs. 2347 Comparison**	**8th Grade *N* = 780 Treatment vs. 2271 Comparison**
Intercept	2643.66*** (46.49)	2645.00*** (53.49)	2637.54*** (54.34)
Treatment (1 = MTP)	-91.02 (60.62)	-85.37 (62.71)	-95.67 (59.94)
Grade 8	3.86 (3.11)	-	-
Year 2019	24.03*** (3.83)	23.85*** (5.62)	23.58*** (5.22)
Year 2018	1.65 (3.74)	2.10 (5.48)	1.20 (5.12)
Gender (1 = Male)	-28.37*** (3.13)	-23.99*** (4.61)	-32.70*** (4.28)
Race/Ethnicity			
American Indian	-61.87***	-66.75***	-57.06***

	(6.25)	(9.24)	(8.44)
Hispanic	-29.86** (11.05)	-31.36 (16.43)	-29.77* (14.82)
African American	-68.03 (34118)	-88.10 (50091)	-48.16 (46286)
Asian	2.76 (76279)	-2.20 (62277)	7.71 (69617)
Other	-37.06** (11.75)	-36.96** (16.94)	-36.30* (16.30)
Special education (1 = Yes)	-186.32*** (5.82)	-197.45*** (8.43)	-175.05*** (8.05)
Economically disadvantaged (1 = Yes)	-17.82 (25.42)	-25.67 (34.39)	-11.80 (38.05)

* *p* < .05, ** *p* < .01

Research Question 4: Sample and Results

Data and Data Cleaning for ACT Outcomes Data

Staff from OPI provided SRI with deidentified data on student ACT scores for the 2016/17, 2017/18, 2018/19 school years. Students in Montana schools take the ACT only during their 11th grade year; thus, we combined the across all years in order to increase the statistical power for our analyses.

Table 11 Sample Sizes for ACT Analyses

Variable	Year	Sample (Grade 11)	Sample (Grade 11 with no missing ACT English and demographic data)
MTP treatment school	Total	587	527
Anaconda High School	2019	82	64
	2018	76	76
	2017	64	64
Browning High School	2019	160	118
	2018	112	112
	2017	93	93
Comparison schools	Total	677	643
Hardin High School	2019	132	113
	2018	100	100
	2017	90	90
Libby High School	2019	70	89
	2018	75	81
	2017	81	85

Wolf Point High School	2019	46	35
	2018	39	39
	2017	34	34
Total		1264	1170

Note. All 10th graders were deleted from the final analytic sample.

Findings Research Question 4

Again, the treatment schools had larger percentages of students who identify as American Indian or were economically disadvantaged than in the control schools. Using the propensity score weighting approach, we were able equate the two groups on these two demographic characteristics and the weighted comparison group.

Table 12 Descriptive Analysis of the ACT Analytic Sample Before Propensity Score Weighting and After Propensity Score Weighting

Variable	**Treatment school (*N* = 527)**		**Comparison Schools before propensity score weighting (*N* = 643)**		**Cohen's d size difference before propensity score weighting**	**Comparison Schools after propensity score weighting (*N* = 643)**		**Cohen's d size difference after propensity score weighting**
Predictors	Mean	*SD*	Mean	*SD*		Mean	*SD*	
School Year 2019	0.35	0.48	0.33	0.47	0.04	0.35	0.43	0.00
School Year 2018	0.36	0.48	0.35	0.48	0.02	0.36	0.43	0.00
School Year 2017	0.30	0.46	0.32	0.47	-0.04	0.29	0.37	0.02
Gender (1 = male)	0.51	0.50	0.54	0.50	-0.06	0.52	0.41	-0.02
Ethnicity								
Hispanic	0.006	0.08	0.05	0.21	-0.33	0.005	0.07	0.01
African American	0.004	0.06	0.003	0.06	0.02	0.004	0.06	0.00

Asian	0.004	0.06	0.008	0.09	-0.06	0.004	0.06	0.00
American Indian	0.61	0.49	0.38	0.48	0.47	0.61	0.44	0.00
Other	0.006	0.08	0.09	0.29	-0.49	0.006	0.07	0.00
Special education (1 = Yes)	0.09	0.28	0.05	0.22	0.15	0.08	0.25	0.04
Economically disadvantaged (1 = Yes)	0.79	0.41	0.79	0.41	0.00	0.79	0.37	0.00
Outcomes								
ACT Reading Scale Score	16.50	5.14	17.74	5.85	-0.23	17.03	5.07	-0.10
ACT English Scale Score	13.84	4.52	15.56	5.17	-0.36	14.82	4.55	-0.21

We again do not observe evidence of an effect of the intervention (Table 13). High school students in treatment schools did not have a statistically significant difference in ACT Reading or English scores from those of students in comparison schools when controlling for demographic characteristics in the propensity score weighted analyses.

As with the SBAC analyses, the lack of a baseline measure of student achievement for analysis makes it difficult for us to determine with any reasonable degree of certainty whether the results of these analyses represent a true null association between participation in the MTP treatment and student reading or English achievement.

Table 13 Propensity Score Weighted HLM Results on ACT Reading Scale Score and English Scale Score, 11th Grade

Variable	**ACT Reading** ***N* = 587 Treatment vs. 677 Comparison**	**ACT English** ***N* = 527 Treatment vs. 643 Comparison**
Intercept	20.27** (0.54)	18.44*** (0.44)
Treatment (1 = MTP)	-0.53 (0.47)	-0.96 (0.36)

Year 2019	0.23 (0.36)	0.04 (0.31)
Year 2018	0.39 (0.36)	0.06 (0.31)
Gender (1 = Male)	-0.56 (0.30)	-0.66** (0.25)
Ethnicity		
American Indian	-3.18*** (0.46)	-3.59*** (0.32)
Hispanic	0.33 (1.98)	-0.47 (1.67)
African American	-0.66 (2.32)	0.85 (1.99)
Asian	-2.06 (2.34)	-1.60 (2.00)
Other	-1.40 (1.93)	-1.45 (1.65)
Special education (1 = Yes)	-3.75*** (0.53)	-2.96** (0.45)
Economically disadvantaged (1 = Yes)	-1.11** (0.47)	-1.07** (0.40)

* *p* < .05, ** *p* < .01

Research Questions 5 and 6: Sample and Results

School leaders in the treatment and comparison schools reported on each school's allocation of literacy instruction time for students in Tiers 1, 2, and 3 via the spring online implementation survey during the 2016–17, 2017–18, and 2018–19 school years. Respondents were the principals of the treatment schools (*N* = 11)[4] and comparison schools (*N* = 8).[5]

On average, across the 3 years of implementation, school principals at treatment and comparison schools reported allocating similar amounts of time for core reading/literacy instruction for students in Tiers 1, 2, and 3. However, principals at treatment schools reported allocating more time than comparison school principals for reading intervention for students performing slightly below grade level (Tier 2) (32.5 minutes versus 27.6 minutes per day) and students performing far below grade level (Tier 3) (51.8 minutes versus 39.6 minutes per day) In addition, principals at intervention schools reported allocating more time than comparison school principals for a replacement core curriculum for Tier 3 students (46.1 minutes versus 35.6 minutes per day).

[4] In Year 4, one treatment school had two administrators respond to the survey questions but with differing responses. Therefore, we have excluded responses for this treatment school from our analyses.

[5] N's represent the total number of survey responses across the 3 years of implementation. The same principal may have completed the survey across multiple years.

Table 14: Principal-reported Average Number of Minutes Per Day Allocated for Tiers 1, 2, and 3 Literacy Instruction, Weighted Average Across Years 2–4

	Treatment Schools		Comparison Schools	
	N	**Mean**	***N***	**Mean**
Core reading/literacy instruction				
Tier 1 students	11	43.9	7	46.7
Tier 2 students	11	47.9	8	48.7
Tier 3 students	11	54	8	53.8
Reading intervention				
Tier 2 students	10	32.5	8	27.6
Tier 3 students	11	51.8	8	39.6
Replacement core curriculum for Tier 3 students	8	46.1	5	35.6

Research Questions 7, 8, 9, and 10: Sample and Results

Teachers reported their literacy instructional practices on an online implementation survey. Reading interventionists and content area teachers were surveyed separately. Among the reading interventionists, across the 3 years of implementation, 15 respondents were from treatment schools and 12 were from comparison schools.[6] More comparison school reading interventionists (67%) than treatment school reading interventionists (27%) reported that their primary role was reading specialist. Of the content area teachers, across the 3 years of implementation, 36 were from treatment schools and 31 were from comparison schools.[7] In both treatment and comparison schools, English language arts (ELA) teachers constituted a large percentage of respondents (treatment schools, 36%; comparison schools, 51%) (Table 15).

Table 15 Primary Role of Teacher Survey Respondents, Weighted Average Across Years 2–4

	Percentage of Respondents in Treatment Schools	Percentage of Respondents in Comparison Schools
Reading interventionist	*n* = 15	*n* = 12

[6] Please note that a reading interventionist may have completed the survey more than once across years 2-4.

[7] Please note that content area teachers may have completed the survey more than once across years 2-4.

General education teacher	60	33
Special education teacher	34	42
Reading intervention teachers/specialist	27	67
English learner teacher	7	0
Content area teachers	*n* = 36	*n* = 31[8]
English language arts teacher	36	51
Math teacher	38	29
Science teacher	19	19
Social studies teacher	5	0
Other (e.g., AVID Elective, supplemental, computer applications, special education, digital literacy and computer science)	14	13

Note. Percentages do not sum to 100% because respondents may have more than one primary role.

Across the 3 years of implementation, teachers in the treatment schools had on average fewer years of experience relative to comparison teachers and were less likely to have relevant qualifications or assistance in their classrooms from other adults. On average, teachers at treatment schools had about half as many years of experience (content area teachers, 11.0 years; reading interventionists, 10.4 years) than those teachers at comparison schools (content area teachers, 19.8 years; reading intervention teachers, 24.0 years). Further, fewer reading interventionists had an endorsement in literacy at treatment schools (7% fewer than those teachers at comparison schools—42%) However, more reading interventionists at treatment schools (40% than at comparison schools—25%) had endorsements in ELA.

Across the 3 years of implementation, teachers at treatment schools reported receiving support from other adults less frequently than did teachers at comparison schools. Although most teachers in treatment and comparison schools reported that they never or rarely have other adults in their classroom, fewer content area teachers at treatment schools reported that they sometimes, often, or always have outside help (17%) compared with comparison school teachers (27%). Across the 3 years of implementation, the majority of reading interventionists in both treatment and comparison schools reported that they provide reading intervention supports and strategies to Tier 2 and Tier 3 students. A higher percentage of reading interventionists at comparison schools than treatment schools reported providing support to Tier 2 students (92% and 80%, respectively). In contrast, all treatment school reading interventionists reported providing support to Tier 3 students, whereas only 75% of reading interventionists at comparison schools reported they did so.

[8] The number of teachers varies for each item, as teachers may have indicated an item was not relevant for their practice (e.g., they did not teach Tier 2 or 3 students in any classes).

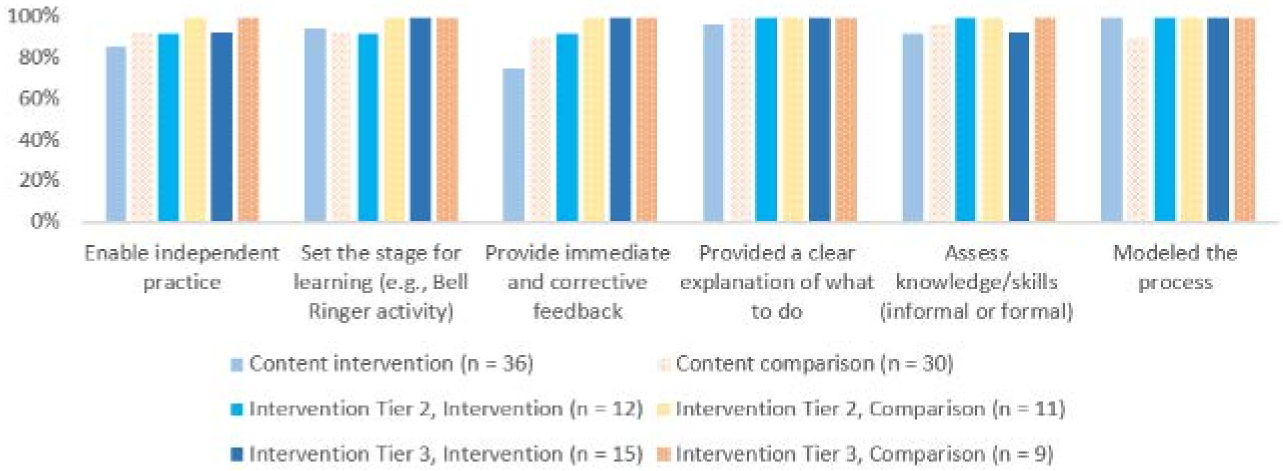
Effective Literacy Instruction and Interventions and Classroom Instructional Support

In general, a majority of teachers at both treatment and comparison schools reported using an explicit instructional approach and implementing literacy strategies to support students in reading and writing. [9]

Explicit instruction

Across the 3 years of implementation, a majority of content area teachers at both treatment and comparison schools (75–100%) reported using several explicit reading activities often or always. Nearly all content area teachers and reading interventionists at both treatment and comparison schools (97–100%) said that they often or always provide a clear explanation of what to do (Figure 2). Of the content area teachers at treatment schools, a lower percentage reported often or always enabling independent practice (86%) and providing immediate corrective feedback (75%) in comparison to the other activities (which ranged from 92–100%).

Figure 2 Teachers' Report of Explicit Instructional Practices That They Often or Always Use in Class During Reading Across Years 2–4



Literacy strategies

On average, across the 3 years of implementation, a higher proportion of comparison school teachers (both content area and reading specialists) tended to report using various types of literacy strategies often or always as compared with treatment school content area teachers and reading specialists. However, the few cases in which treatment school teachers were more likely to report using a specific literacy strategy occurred for strategies that were components of the intervention. For example, a greater percentage of treatment school content area teachers

[9] Explicit instruction is a structured, systematic, and effective methodology for teaching academic skills. During reading and writing activities, it is characterized by a series of literacy supports or practices whereby students are guided through the learning process.

and reading specialists teaching Tier 2 students reported often or always using AVID strategies and the LINC vocabulary strategy when compares to comparison school teachers. In addition, among reading interventionists teaching Tier 2 and 3 students, a greater percentage at comparison schools reported that they often or always use the Unit Organizer Routine. Further, treatment school reading interventionists targeting Tier 3 students were more likely than their counterparts at comparison schools to report using the Question Exploration Routine (QER) often or always (Table 16).

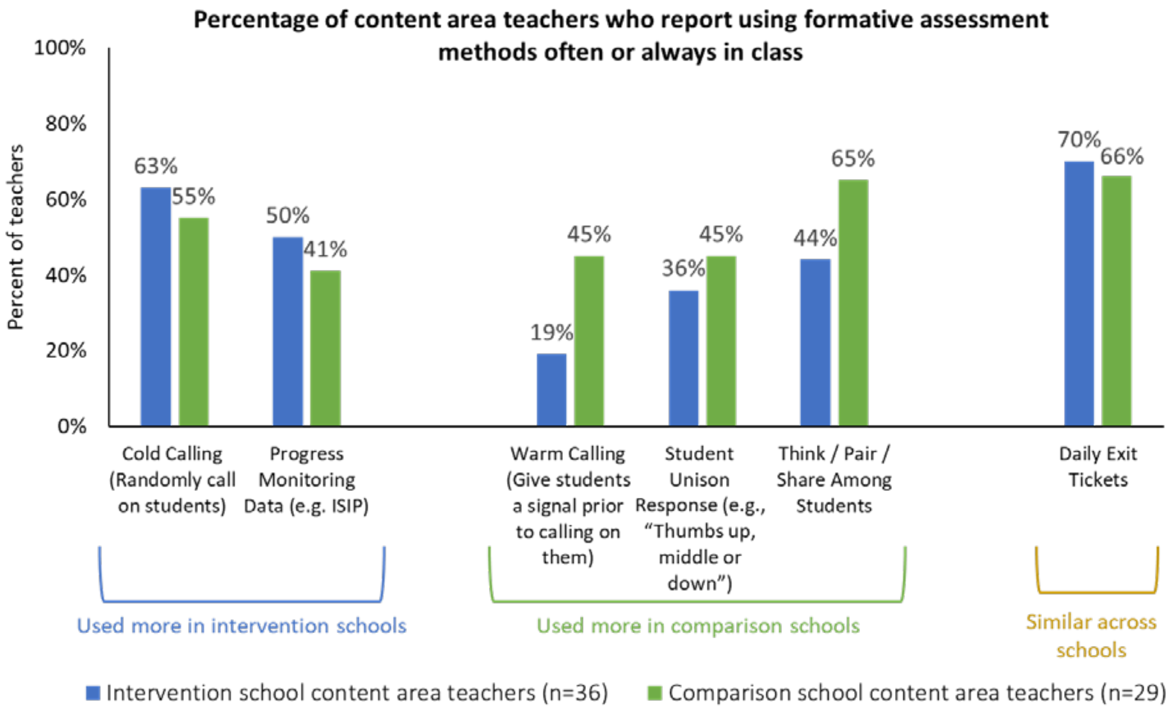
Table 16 Teachers Who Report They Often or Always Use Literacy Strategies

	Content area teachers		Reading intervention teachers: Tier 2		Reading intervention teachers: Tier 3	
	Intervention schools (n = 36)	Comparison schools (n = 29)	Intervention schools (n = 12)	Comparison schools (n = 11)	Intervention schools (n = 15)	Comparison schools (n = 10)
MTP intervention strategies						
Course Organizer Routine (SIM strategy)	14	41	8	18	14	23
Unit Organizer Routine (SIM strategy)	47	38	41	27	53	33
Question Exploration Routine (SIM strategy)	30	45	42	45	60	33
AVID strategies (e.g., Cornell Notes, Learning Log, Cornell Notes Rubric)	69	53	92	55	80	78
Overall vocabulary intervention strategies						
LINC Vocabulary Strategy (SIM strategy)	53	22	58	36	47	45
Explicit vocabulary strategy	53	48	83	91	80	100
CRISS Strategies	29	55	17	36	33	75
Non-MTP intervention-specific strategies						
Lesson Organizer Routine (SIM strategy)	22	27	25	36	33	45
Explicit prediction strategy	28	31	17	63	27	88
Explicit summarization strategy	45	38	42	55	47	78
Graphic organizers (e.g., compare / contrast, main and supporting ideas)	53	61	50	100	40	80
Use of context clues	50	82	92	100	80	90
Rehearsing information aloud	56	59	92	100	86	100
Mnemonic devise for remembering information	25	41	33	64	60	100

Use of formative assessments in instructional decision-making

Across Project Years 2, 3, and 4, treatment and comparison school content area teachers differed in their reported formative assessment methods. Among content area teachers, a higher percentage at treatment schools than at comparison schools reported often or always using cold calling and progress monitoring data. In contrast, more comparison school content area teachers reported that they often or always use warm calling, student unison response, and think/pair/share. A similar percentage of treatment and comparison school content area teachers reported using daily exit tickets often or always in class (Figure 3).

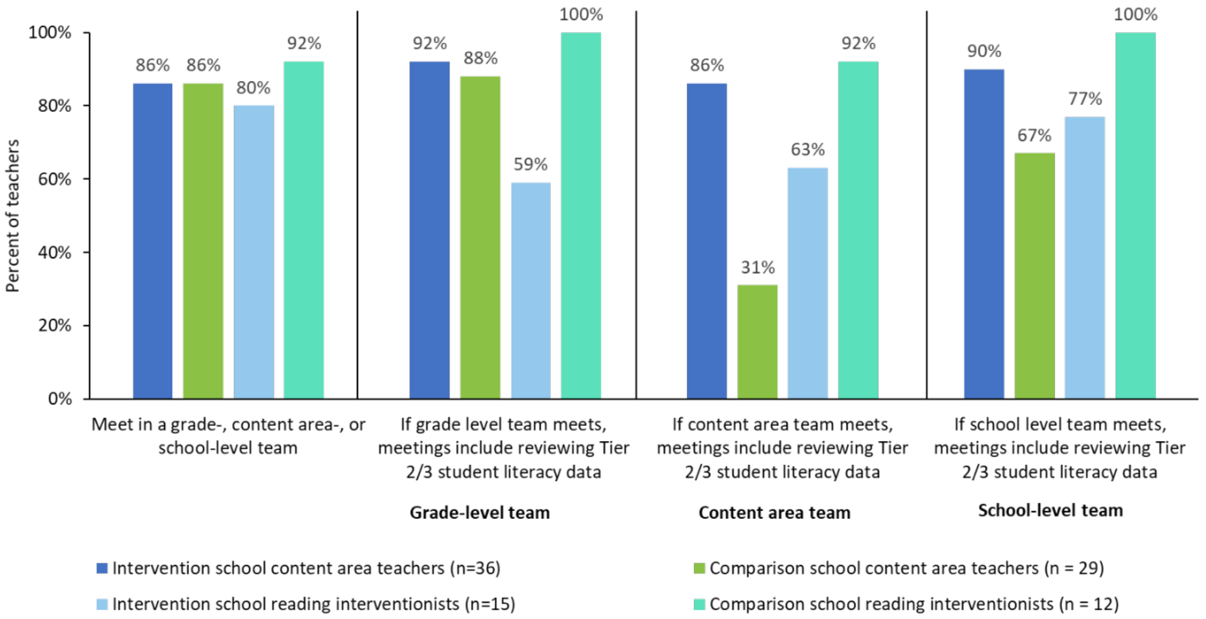
Figure 3 Content Area Teachers Who Report Using Formative Assessment Methods Often or Always in Class



Note. For the treatment school teacher sample, we took the weighted average of years 2, 3, and 4. For the comparison school teacher sample, this only includes year 2 and year 3 respondents because no comparison school content area teachers completed the survey in year 4.

higher percentage of treatment teachers reported reviewing these data in content area teams (86% vs. 31%) and school-level teams (90% vs. 67%). However, a lower proportion of treatment school reading interventionists as compared with comparison school reading interventionists reported participating in grade, content area, and/or school teams to monitor student progress (80% and 92%, respectively), and discussing Tier 2 and Tier 3 student literacy data in grade-level teams (59% vs. 100%), content area teams (63% vs. 92%), and school-level teams (77% vs. 100%).

Figure 4 Content Area and Reading Intervention Teachers' Report of Meeting in Grade, Content Area, or School Teams to Monitor Student Progress



Conclusion

Overall, we do not find an association between participation in the Plan-Do-Study-Act treatment on schoolwide student outcomes within or across the 3 years of implementation. This does not mean that the project was failure. First, we should interpret these findings with caution. Not all students in the intervention schools received the treatment and, even with matching, intervention and comparison schools were different in important ways. In addition, without baseline measures and longitudinal data following each student, it is difficult to accurately capture improvements in student outcomes even with our matching procedure and use of covariates.

The PDSA change ideas implemented in each school were small, in order to be a manageable adjustment that all teachers could incorporate into their classroom routines during the semester. In addition, these changes were made at two time points within the school year for a brief period of time. There was no expectation that researchers that teachers would continue using the change ideas following the PDSA cycle conclusion; however, survey results show treatment teachers did continue to implement these strategies after the conclusion of the PDSA cycle.

Although the analyses did not identify impacts on students, participating teachers noted that the process was helpful and that they valued the opportunity to connect with one another and discuss their practice. The PDSA implementation guidebook, created through a collaboration between SRI, OPI, and the teachers and administrators in participating schools, now provides teachers and educators across Montana with practical information on setting up and executing a PDSA cycle in their own schools.

References

ACT. (2019). *ACT technical manual.* ACT, Inc.

American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Authors.

Becker, S. O., & Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal, 2*(4), 358-377.

Curtis, L. H., Hammill, B. G., Eisenstein, E. L., Kramer, J. M., & Anstrom, K. J. (2007). Using inverse probability-weighted estimators in comparative effectiveness analyses with observational databases. *Medical Care, 45*(10), S103-S107.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15*(3), 234.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica, 71*(4), 1161-1189.

Mather, N., & Woodcock, R. W. (2001). *Woodcock-Johnson III tests of achievement: Examiner's manual*. Riverside Publishing.

Mathes, P. (2016). *Istation Indicators of Progress (ISIP) Advanced Reading technical report: Computer adaptive testing system for continuous progress monitoring of reading growth for students grade 4 through grade 8.* Imagination Station, Inc.

Noble, J., Davenport, M., Schiel, J., & Pommerich, M. (1999). *High school academic and noncognitive variables related to the ACT scores of racial/ethnic and gender groups* (ACT Research Report No. 99-6). ACT.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association, 79*(387), 516-524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33-38.

Smarter Balanced Assessment Consortium. (2016). *2013–2014 Technical report: Validity, item and test development, pilot test and field test, and achievement level setting*. Authors.