SRI Education™
A DIVISION OF SRI INTERNATIONAL

Evaluation of National Writing Project's College-Ready Writer's Program 2015 SEED Grant

Technical Report

2019

SRI International · 333 Ravenswood Avenue · Menlo Park, CA 94025-3493 · phone +1.650.859.2000 · www.sri.com

This report was prepared by SRI International with funds provided by the National Writing Project under a grant from the U.S. Department of Education. However, the contents do not necessarily represent the policy of the U.S. Department of Education, and you should not assume endorsement by the Federal Government.

Suggested Citation: Arshan, N. L., Park, C. J., & Gallagher, H. A. (2019). *Evaluation of National Writing Project's College-Ready Writer's Program 2015 SEED grant: Technical report.* Menlo Park, CA: SRI International.

Contents

Contents

Exhibits

Introduction

The National Writing Project's (NWP) College, Career, and Community Writers Program (C3WP) provides professional development for teaches in grades 7–10 with the goal of improving students' source-based argument writing. C3WP aims to build teachers' understanding of and skill in teaching argument writing. C3WP was subject to a prior rigorous evaluation, which found that two years of C3WP had a positive impact on writing outcomes for students in grades 7-10 in high-need rural districts (Gallagher, Arshan, & Woodworth, 2017). C3WP was formerly called the College-Ready Writers Program (CRWP).

In 2015, with the support of a Supporting Effective Educator Development (SEED) grant, the National Writing Project sought to extend the reach of the program to non-rural contexts and to examine the efficacy of a short-cycle version of the program. The SEED grant also supported researchers at SRI Education to conduct an independent evaluation (a within-teacher randomized controlled trial) designed to examine both program implementation and the impact of the program on teachers' instructional practices and student learning. This report presents the results of the short-cycle impact analysis that focused on 7th and 8th grade English Language Arts (ELA) teachers in high-needs districts. "High needs" is defined as school populations where at least 50% of students are eligible for free or reduced lunch programs.

The technical report begins with a discussion of the C3WP program components and intended outcomes. Then, it describes the research design including a discussion of recruitment and randomization; site, teacher, class, and student samples; and data and methods. Finally, the report provides findings related to program implementation, teacher practice outcomes, and student learning outcomes.

C3WP Program Design

C3WP is designed to change teachers' understanding of and instructional practices in source-based argument writing, with an ultimate goal of helping students to become skilled at writing arguments based on information in non-fiction sources. As noted above, a prior evaluation of C3WP examined the impact of a two-year version of the program on students' source-based argument writing. This evaluation examines a short-cycle version of the program as implemented by six local Writing Project programs.

C3WP has three core features (1) intensive professional development; (2) instructional resources; and (3) formative assessment, described below.

- **Professional development.** C3WP calls for intensive and embedded teacher-to-teacher professional development focused on supporting classroom implementation of argument-writing teaching practices. This includes 45 hours of professional development that uses strategies intended to support implementation such as modeling lessons, co-teaching, and planning for implementation.
- **Instructional resources.** C3WP includes instructional resources designed to scaffold the teaching of argument writing for teachers by providing models which teachers can learn from and use in the classroom. Each resource focuses on a specific skill or practice in argument writing, and the resources are designed to build on each other. Local Writing Project sites were expected to support teachers in learning and using the instructional resources in their

classrooms. Teachers were expected to implement four cycles of argument writing using the instructional resources.

- **Formative assessment.** C3WP features regular formative assessment to inform subsequent steps in instruction. The Using Sources Tool (UST), a C3WP-specific formative assessment resource, supports teachers in analyzing how students make and support claims using evidence from sources. Teachers were expected to use the UST at least twice over the course of the year to examine their students' writing. Sites were expected to review the UST in professional development sessions and collaboratively examine the data.

Local university-based writing-project sites provided C3WP professional development and adapted the program to local context while maintaining the core features. NWP's national office provided technical assistance to local sites in the form of national convenings of sites intended to build the sites' and teachers' understandings of C3WP, connect with and learn from other local sites, and reflect on their work together. The program began with a launch in summer 2016 and a mid-year convening in February 2017. The national office also maintained a website with C3WP materials and resources and convened periodic virtual meetings.

Student learning outcomes

C3WP aims to improve students' abilities to write an argument using evidence from non-fiction sources. The program introduces students to specific argument writing skills such as identifying and responding to arguments; selecting and annotating evidence; crafting a claim with evidence; and organizing evidence. The expectation is that introducing these skills will improve students' performance on a source-based argument writing task. More broadly, C3WP seeks to "support students in reading critically, exploring multiple points of view, and taking a stand on important issues" (Friedrich, Bear, & Fox, 2018, p.19).

Teacher instructional practices

C3WP asks teachers to implement cycles of instruction aimed at developing students' skills in argument writing using C3WP instructional resources and to engage students in routine argument writing (RAW). Within the context of C3WP, a cycle of argument writing starts with teachers experiencing an instructional resource in professional development, trying the instructional resource in the classroom, and examining student work collaboratively during professional development using C3WP's formative assessment tools to determine instructional next steps. The participating 7th and 8th grade ELA teachers were asked to teach four cycles of C3WP argument writing instruction during the year and analyze their students' argument writing with the UST at least twice.

State, district, and school context

The SEED grant required that participating teachers had to be teaching within schools where at least 50% of the student populations qualified for free and/or reduced-price lunch. In addition, all of the schools in our study were located in states which had adopted the Common Core English Language Arts standards, which featured a shift towards argument writing. During recruitment, district staff consistently shared that teachers' writing instruction was already Common Core aligned.

Research Design

The evaluation was a classroom-randomized, controlled trial measuring the impact of C3WP on teachers' 7th and 8th-grade ELA instructional practices and students' source-based argument writing achievement. Our primary research questions include:

1. Was C3WP implemented with fidelity?
2. What impact did C3WP have on teacher practice?
3. What impact did C3WP have on student source-based writing achievement?

Recruitment, randomization, and the counterfactual condition

NWP invited a subset of their local Writing Project sites to participate in this evaluation. The sites participating in the study each had prior experience with C3WP and/or a history of conducting in-service professional development in schools. Each site recruited schools in their service area to partner with on the program. Within these schools, sites recruited 7th through 8th grade ELA teachers to participate in the experiment.

Prior to randomization, the research team asked teachers to nominate two similar classrooms: one would be randomized into C3WP, the other into a business-as-usual control group. To participate in the study, teachers committed to participate in C3WP training and to use the tools and materials in their C3WP classroom. They also committed to wait until after post-test to use any C3WP tools, materials, or, as possible, strategies, in the control classroom. Student pre-test writing skills were assessed within three weeks of the first day of school. Teachers administered the post-test during a four-week period from mid-February to mid-March. The timing of the pre- and post-test writing assessments allowed researchers to collect student rosters from stable classrooms prior to randomization and allowed teachers time to implement a C3WP unit prior to state testing. On average, this timing allowed for a 6-month window to implement C3WP.

A within-teacher randomization of a program designed to improve teacher effectiveness comes with limitations, risks, and benefits. First, as the teacher is constant between the treatment and control conditions, any general skill developed by the teacher that would benefit the students' source-based argument writing without using C3WP strategies, tools, and materials (e.g., greater understanding of argumentation) will not be included in the estimated treatment effect. As such, the estimated impacts of this design may underestimate the full impacts of C3WP and should be interpreted as estimating only the impacts of the programs' strategies, tools, and materials.

Second, the risks of contamination are high, as teachers have access to program strategies, tools, and materials and could use them in the control classrooms. To minimize this risk, local Writing Projects recruited schools and, within those schools, gave teachers the option to participate in the research study or opt out. Further, the student outcome measure was collected at least 6 weeks before state standardized testing so that teachers would have a chance to use C3WP in the control classrooms prior to any high-stakes assessment, thus mitigating the risk that assessment pressures might lead teachers to use C3WP in their control classrooms. Teachers who chose to participate in the study were screened by the research team for understanding of and buy-in for the research design. These teachers received a stipend of $1,000 for participation to provide a further incentive to maintain the distinction between the treatment and control

conditions. The research team monitored contamination through daily instructional logs and end-of-year interviews (described below).

Despite these limitations and risks, the design also provides a strong benefit of cost-effectiveness. A study of 62 classrooms can be executed for far less money than a study of 62 schools and does not require the disruptions of communities of practice inherent in a study that randomizes teachers. Second, the within-teacher design controls for all other aspects of teacher effectiveness except the use of C3WP tools and materials, likely improving overall statistical power (Rhodes, 2011).

Study samples

Writing Project site and district samples

NWP recruited six local Writing Project sites with an history with providing in-service development to schools and districts. Five of the six sites had implemented C3WP in the prior evaluation. The leadership of the sixth site had been involved in the leadership and development of C3WP. The local Writing Project sites recruited high-needs schools serving grades 7 and 8 in their service areas. Four of the sites worked with teachers across two schools, and two sites worked with teachers within one school.

Teacher samples

Local Writing Project sites worked with their partner schools to identify eligible teachers and target them for recruitment. To be eligible for inclusion, teachers must have been a core ELA teacher, teach at least two eligible ELA classes in grades 7 and/or 8, and make a 2-year commitment to the study. In addition, teachers agreed to implement C3WP in their focal class and not use C3WP materials in their control class. The teacher sample at randomization consisted of 31[1] 7th and 8th grade ELA teachers (randomization blocks) and 62 classrooms (units of randomization).

We calculate attrition as

$$Attrition = \frac{N_{assigned\_sample} - N_{analytic\_sample}}{N_{assigned\_sample}}$$

One teacher attrited from the study after randomization, removing two study classrooms (one C3WP and one control) from the analytic sample. Cluster-level attrition is therefore 3.2% overall and for each study condition.

Class and student samples

The research team asked teachers to choose classes that were the in same grade level, contained the same target population (e.g., both regular track or both honors), and had similar achievement and behavior patterns (to the best of their ability to observe within the first few weeks of school). Researchers reviewed the teachers' class selections and collected student rosters for the selected classes prior to randomization.

---

[1] The originally randomized sample consisted of 32 teachers and 64 classrooms. One teacher left the study prior to collection of baseline data (and therefore notification of which classroom had been randomized into each condition). As such, we report on a randomized sample of 31 teachers.

Before randomization, the research team collected rosters of students enrolled in each teacher's study classes. Students enrolled in these study classes were randomized as a cluster. The research team asked schools to maintain study classes intact absent unavoidable mitigating circumstances. No students joining a study classroom after collection of student rosters were included in the analytic sample.

- If students were moved to a classroom in the alternate treatment condition, they were analyzed according to their originally assigned status and teacher (an intent-to-treat estimate).
- As the student outcome data were researcher-collected measures (see below for more detail), the research team was unable to collect data if a student left the study classroom.
- Students were considered attrition if they transferred schools, transferred into a non-study classroom, or were absent for or declined to participate in data collection.

Student attrition was 23% overall, and for each group (Exhibit 2).

To examine cross contamination of classes with C3WP practices, we collected instructional-practice data from logs (described below), and researchers conducted interviews with 13 teachers. Minimal evidence of contamination was reported (e.g., one teacher searched for a graphic organizer similar to one provided by C3WP for use in her control classroom).

Exhibit 1. Student Attrition by Treatment Status

	Treatment	Control	Overall
Assigned *n*	614	605	1,219
Analytic *n*	474	465	939
Attrition	23%	23%	23%

NOTE: Assigned *n* excludes students whose teacher attrited from the study.

Data and methods

The research team collected data from multiple sources to understand program implementation and assess outcomes. We examined program implementation across the six Writing Project sites, measured teacher practice through a daily teacher instructional log, and assessed student learning with a study-administered, on-demand student writing assessment.[2]

Program implementation

The research team worked with NWP to develop indicators for each of C3WP's core components to assess whether the program was implemented with fidelity across the local Writing Project sites. As described above, the three key components are:

- Intensive and job-embedded professional development
- Support for the use of C3WP instructional resources
- Use of formative assessment to examine student work

[2] The research team also interviewed professional development leads and teachers at each site but does not draw on the interviews to describe implementation in this technical report.

To assess the fidelity of implementation (FOI), the research team obtained data on the extent of teacher participation in the professional development, features of the professional development program, and use of the UST to conduct formative assessment. Each measure was examined against thresholds for each indicator to assess whether the site fully implemented the program. We specify how each feature was operationalized in the context of this one-year version of the program. Exhibit 2 lays out each indicator and threshold.

To assess implementation, researchers obtained administrative data from NWP's Professional Learning Tracker (PLT) and UST online platform. Each local Writing Project site was required to submit PLT data three times a year on the participants, hours, content, and strategies associated with each professional development event. The researchers used this data to determine the extent to which the sites implemented the program as intended.

Exhibit 2. C3WP Implementation Measures

	Key Elements of Feature	Operational Definition for Indicator	Data Source(s) For Measuring Indicator	Teacher-level Threshold	Site-level Threshold
Component 1: Duration and Breadth of Teacher Professional Development					
1.1	Duration and breadth of participation	80% of participating 7-8 ELA teachers participate in 45 or more hours of PD	PLT	A: 35+ hours B: 40+ hours	85%+ of participating 7–8 grade ELA teachers reach teacher-level threshold A OR 75% of participating 7-8 grade ELA teachers reach teacher-level threshold B
Component 2: Content of the Professional Development					
2.1	Focus on argument writing	80% of participating 7–8 grade ELA teachers participate in 36 hours of PD focused on argument	PLT	A: 25+ hours B: 30+ hours	85%+ of participating 7–8 grade ELA teachers reach teacher-level threshold A OR 75%+ of participating 7–8 grade ELA teachers reach teacher-level threshold B
2.2b	Use of formative assessment tool	Analysis of student work is submitted via NWP's Using Sources online tool on 2 occasions	NWP's Using Sources Tool data base online platform		Analysis of student work is submitted via NWP's Using Sources online tool on 2 occasions prior to end of year (or outcome administration)
Component 3: Professional Development Strategies					
3.1	Focus on classroom enactment	For 80% of participating 7-8 grade ELA teachers, 50% of PD events focus on classroom enactment, including: (1) demo lesson, 2) designing tasks/assignments, (3) planning for classroom implementation, (4) analyzing student work, (5) analyzing student work, (6) Modeling instruction with teachers, (7) co-teaching/co-planning; and (8) debriefing classroom implementation	PLT	50% of PD events	75%+ of participating 7–8 grade ELA teachers reach teacher-level threshold

Teacher practice outcomes

The measurement of teacher practice outcomes was based on data from an instructional log. The data source and impact analysis are described below.

Data

We measured teacher practice through an instructional log, administered to teachers daily for two weeks during the study. This instrument was adapted by the research team from a similar instrument found to have high rates of inter-rater reliability (Gallagher, Arshan, & Woodworth, 2017). The research team adapted the logs to be appropriate for a within teacher comparison by focusing on instruction closely aligned to C3WP's strategies, tools, and materials. The teachers answered each set of questions twice: once for the C3WP classroom and once for the control classroom. To monitor contamination between the classrooms, the log also asked about the use of C3WP tools and activities in each class. The log was closely aligned to the program's goals, with a focus on time spent writing and the skills students practiced during this writing time. To measure impacts on instruction, the log included questions about instructional time spent writing, the purposes for writing, and the nature of the students' revisions or argument writing, if one of those activities took place.

The logs were administered daily to more accurately capture the specific classroom practices enacted, and they were sampled on multiple days to minimize the error associated with the measurement of teacher practice. To improve the precision of the estimates, the log prompted teachers to focus on an average individual student in the class rather than an estimate across students.

Impact estimates

To estimate impacts on practice, we used a multi-level model, in which the predicted practice Y on day i of class j taught by teacher k is given as:

$$\widehat{Y_{ijk}} = \beta_0 + \beta_1(C3WP_i) + \sigma_k + \delta_j + \eta_i.$$

Random effects σ_k, δ_j, and η_i account for error at the teacher, class, and day levels. Typically, teacher-fixed effects would be a more appropriate modeling choice, as variance in treatment effects by teacher combined with an unequal sample size within classes might bias estimates. However, teachers who only answered a question for one condition or the other (e.g., those who only taught writing in their C3WP class during log administration) and therefore have missing data describing their writing instruction would not contribute to the estimated treatment effects using teacher-fixed effects models. As such, we run random effects models and used a teacher-fixed effect model as a specification check to ensure results are not driven by bias caused by variation in teacher effects. Logistic models were estimated using Stata version 14.2's *meqrlogit* command; the one continuous outcome used the *mixed* command. All logistic outcomes are converted into predicted percentage of days for ease of interpretation.

Teachers used C3WP tools and materials so infrequently in control classrooms that multi-level logistical models would not converge with consistency when estimating these differences by treatment condition. We therefore provide descriptive statistics (i.e., means and sample sizes) on the use of C3WP tools and materials by condition to describe any possible contamination.

Student outcomes

For this study, student outcomes were measured using a source-based argument writing assessment. The data used and analysis are described below.

Data

The research team administered a two-day writing assessment to all students in study classrooms in fall 2016 (baseline) and again in spring 2017 (outcome). On day 1, the prompts provide students with approximately five sources on one of four topics. Source material included articles, images, graphs, and infographics. The topics were selected to be of interest to students and provide multiple potential arguments (e.g., payments for college athletes). On day two, the outcome prompt asked students to write an argument using evidence from the source materials. Teachers administered the baseline prompt before learning of classes' randomization status and received copies of the baseline writing for their C3WP classrooms' writing to allow students to use the sample for revision.

To account for any prompt effects (i.e., one topic or set of writings on which students would tend to score higher or lower than average) and to avoid over-alignment of the outcome measure to the intervention (i.e., teachers or students having access to the prompt used at outcome assessment), prompts were counterbalanced so that all classrooms within a local Writing Project site were assigned one prompt in the fall and the other in the spring. This design meant that (1) prompt effects were balanced across local Writing Project sites, and (2) no study teacher would have access to the post-test prompt their students would write about until the day of the post-test assessment.

The student writing was scored with the Analytic Writing Continuum for Source-Based Argument (AWC-SBA). Over a decade, NWP developed the Analytic Writing Continuum (AWC), which has been shown to be a valid and reliable measure of student writing (Bang, 2013).[3] The original version of the AWC had been used primarily to score writing based on students' personal experience and did not explicitly measure the use of evidence from other sources. To assess source-based arguments, NWP developed writing prompts that would require students to select and use evidence from written sources to support their claims or to inform an audience about a particular issue. The resulting performance tasks are similar to the performance-based tasks that were part of some state assessments (e.g., Connecticut) and are part of national assessment consortia (i.e., PARCC and Smarter Balanced). NWP worked with a panel of writing assessment experts to modify the AWC to more accurately score writing that relied on the use of external sources as evidence. The same panel of writing assessment experts selected and annotated anchor papers to be used in training scorers. The revisions to the AWC and the development of annotated anchor papers were designed to help make explicit how well-established attributes of effective writing are evident in source-based argument writing. For example, the AWC-SBA's rubric for the stance attribute directs reviewers to assess the extent to which the writing establishes the credibility of the cited source material. The resulting AWC-SBA retains a basic structure rooted in the "6+1 Traits" of writing (Culham, 2003) but has a particular focus on the attributes related to source-based argument writing.

[3] For more information about the National Writing Project Analytic Writing Continuum, see http://www.nwp.org/cs/public/print/resource/3776

The AWC-SBA measures four attributes:

- Content (Including Quality of Reasoning and Use of Evidence) describes how effectively the writing presents an argument supported by reasoning and the use of evidence from sources.
- Structure describes how effectively the writing establishes an order and arrangement to enhance the central argument.
- Stance communicates a perspective through tone and style appropriate for the purpose and describes how effectively the writing establishes credibility.
- Conventions describe how effectively the writing demonstrates age-appropriate control of usage, punctuation, spelling, capitalization, and paragraphing.

While C3WP has potential to impact all four attributes measured by the AWC-SBA, the research team designated the content attribute as the best aligned measure and the study's confirmatory contrast, as in the prior study (Gallagher, Arshan, & Woodworth, 2017).

For unbiased administration and scoring, local research site coordinators were hired to support and monitor data collection in person in all schools. These local research coordinators de-identified the samples by removing names or other personal identifying information, then scanned the papers, attaching an anonymized identification number provided by the research team. The research team checked the scanned writing for legibility, completeness, and remaining identifiable information, then sent the de-identified papers to NWP for scoring. Therefore, scorers did not know the district the papers came from, the timepoint of data collection, or the treatment status.

Scorers were recruited from current and former teachers with middle school experience affiliated with local Writing Project sites not participating in C3WP (to limit the potential for bias in scorers familiar with the program). Scorer training include initial self-review of the AWC-SBA rubric, anchor papers, prompts, and support materials, followed by group a day of group training prior to scoring. Scorers participated in intermittent calibrations throughout the scoring. Small groups of scorers were led by table leaders. These table leaders had prior experience scoring using the AWC-SBA and received an additional day of training prior the scorer training. They worked to support scorers at their table by answering questions and "reading behind" scorers (i.e., performing informal checks of scorers' ratings throughout the process). Scoring was conducted online using NWP's AWC-SBA's online scoring system. Scorers logged in remotely and scored during specific blocks of time over the course of several days. Exhibit 3 provides descriptive statistics of student writing achievement on each of the AWC-SBA's attributes by treatment condition and timepoint. Table Leaders reviewed scorings and provided individual support at scheduled times throughout the days while group calibrations occurred at regular intervals along with individual calibrations and additional support as needed.

Exhibit 3. Baseline and Outcome AWC-SBA Scores by Time and Treatment Condition

	Baseline			Outcome		
	Treatment	Control	Overall	Treatment	Control	Overall
Content	2.6	2.5	2.5	3.1	2.8	3.0
(*SD*)	(1.1)	(1.0)	(1.0)	(1.2)	(1.0)	(1.1)
Structure	2.6	2.4	2.5	3.0	2.8	2.9
(*SD*)	(1.0)	(1.0)	(1.0)	(1.2)	(1.0)	(1.1)
Stance	2.7	2.6	2.6	3.2	2.9	3.0
(*SD*)	(1.1)	(1.0)	(1.1)	(1.2)	(1.1)	(1.2)
Conventions	2.7	2.6	2.6	3.1	2.9	3.0
(*SD*)	(1.2)	(1.1)	(1.1)	(1.2)	(1.2)	(1.2)
n	474	465	939	474	465	939

NOTE: Means are unadjusted for differences in grade level or prompt form taken by students.

On the six-point rubric for the content attribute, 80% of papers at outcome received scores from 2 to 4. Papers scoring a 1 (10% of papers for the content attribute) were typically too brief to evaluate, consisted primarily of copied text, or lacked any discernible central idea (e.g., haphazard content). About 7% of the outcome papers received a 5 or a 6 for content. These papers competently (score 5) or effectively (score 6) demonstrated reasoning by selecting and using evidence from sources to support the claim and included commentary on that evidence. Although papers received a different score for each attribute, the scores overall reflect a high degree of internal consistency (see Bang, 2013 for a discussion of internal consistency using the AWC and Gallagher, et. al., 2017 for excerpts of student writing at score points 2–4 from the prior study).

Reliability of the prompt scoring was assessed separately for each writing-attribute measure in the AWC-SBA through the double scoring of a subset of papers. Researchers randomly selected 24% of the papers to be double scored and calculated the percentage of papers for which individual scorers agreed within a score point for each attribute. A total of 442 papers in the analytic sample were double scored; raters agreed within a single score point for 88% of papers on the content and stance attributes, 89% on the structure attribute, and 86% on the conventions attribute.

Student impact estimates

This technical report provides supporting detail to models published elsewhere. The Ordinary Least Squares (OLS) models with classroom-fixed effects, described first, are documented in the research brief, "Impacts on Students of a Short-Cycle Implementation of the National Writing Project's College, Career, and Community Writers Program," (Arshan, Park, & Gallagher, 2018). The Hierarchical Linear Models (HLM) will be documented a forthcoming academic publication. These models are written to address different audiences and publications and, therefore, are not directly comparable as specification checks to each other (although comparison of results from different model estimations provides some robustness checks).

Other published texts and this technical report interpret findings in a manner consistent with impact estimates across all models.

OLS Classroom-Fixed Effects Models. The OLS models are intended to compare directly to an earlier research brief describing the impacts of the two-year version of CRWP (Gallagher,

Woodworth, & Arshan, 2015) and, as such, use raw AWC-SBA scores as outcome data. As these results were intended to be included in a two-page form without room for specification checks, we ran OLS models with classroom-fixed effects to conservatively account for the differences in sizes between classrooms. These models include controls to adjust for differences in grades and prompt forms. This report provides these OLS estimates using standardized outcome scores to allow for comparison of point estimates and standard errors across model specifications. The estimated impacts are relatively small, and the standard deviations of the outcome variable are close to 1, so the point estimates are identical when reporting raw or standardized effect sizes rounded to two decimal places. The standardized effect sizes provided in this report are therefore equivalent to the raw impacts reported in the policy brief this report supports.

The OLS classroom-fixed effects models predict writing ability for student *i*, in classroom *j*, taught by teacher *k* as a function of the classroom's assignment to treatment as:

$$Y_{ijk} = (\boldsymbol{Teacher_k})\boldsymbol{\beta_1} + \left(\boldsymbol{C3WP_j\ x\ Teacher_k}\right)\boldsymbol{\beta_2} + (\boldsymbol{Covar_i})\boldsymbol{\beta_3} + \sigma_{ijk}.$$

The vector of student-level covariates includes the pre-test score for each of the four attributes and a rater-fixed effect to account for variation in ratings by scorer. Student baseline and outcome scores are standardized within cohort and prompt form to account for prior achievement, cohort at baseline, and prompt effects. Individual student error is modeled by σ_{ijk}. This OLS model suppresses a constant term and predicts both teacher-fixed effects and a treatment-by-teacher fixed effect. $\boldsymbol{\beta_2}$ therefore provides individual estimates of the effect of each classroom's assignment to C3WP on student writing performance, standardized within the analytic sample.

To combine these individual effects into an overall effect, we used Stata 14.2's *lincom* command to average these 30 estimates, allowing for 29 degrees of freedom. The resulting estimate therefore accounts for the clustering of students within classrooms and the variation in treatment effect across teachers without relying on the assumption of this variation in a treatment effect having a normal distribution, as with HLM models. Further, the classroom-fixed effects models account for varying classroom sizes within teachers.

HLM Models. The HLM models were written for a research audience. As such, we used HLM models to account for clustering of students within classrooms and classrooms within teachers. HLM models are common within research audiences and allow for two interpretations of the impacts: the teacher-fixed effects model treats the sample teachers as a convenience sample that cannot be generalized to other teachers, and teacher random effects treat the teachers as representative of a larger distribution of teachers. For HLM models, we standardized all AWC-SBA data within prompt and grade-level means to account for differences in cohorts and prompt forms; as such, the impact estimates from these models provide estimated effect sizes and cannot be directly compared to the sizes of impacts estimated by OLS models.

HLM models predict writing ability for student *i*, in classroom *j*, taught by teacher *k* as a function of the classroom's assignment to treatment as:

$$Y_{ijk} = \beta_0 + \beta_1\left(C3WP_j\right) + (\boldsymbol{Covar_i})\boldsymbol{\beta_2} + \left(\boldsymbol{Covar_j}\right)\boldsymbol{\beta_3} + \sigma_{ijk} + \delta_{jk} + \eta_k.$$

As in the OLS classroom-fixed effects models, the vector of student-level covariates includes the pre-test score for each of the four attributes and a rater-fixed effect, to account for variation in ratings by scorer. As these models do not include classroom-fixed effects, we additionally

include a vector of classroom covariates, which adjust for classroom average baseline scores for each of the four attributes measured by the AWC-SBA. A likelihood-ratio test indicates that inclusion of these additional classroom-level covariates improves model fit, $\chi^2(4) = 14.95$, $p = 0.0048$. Student baseline and outcome scores are standardized within cohort and prompt forms to account for prior achievement, cohort at baseline, and prompt effects. Individual student, classroom, and teacher errors are modeled by random effects σ_{ijk}, δ_{jk}, and η_k. Student and classroom errors are estimated as random in all models; we discuss the difference between teacher-fixed and random error terms below. In all models, $\boldsymbol{\beta_1}$ provides an average estimate of the effect of classroom assignment to C3WP on student writing performance, standardized within the analytic sample.

In the teacher-fixed effects models, δ_{jk} provides a fixed effect that accounts for variation between teachers and the blocked random assignment of classrooms within teacher. To the extent that the teachers in this study represent a convenience sample that cannot be generalized to a broader population, this model estimation may be preferable, as effects are averaged across teachers without reliance on an assumption about the distribution of effects at the teacher level. These models therefore guard against any skew that may be created due to outlier teachers or differences in class sizes between teachers.

A reasonable case could be made, however, that these teachers represent a population of middle-grades teachers at high-needs schools whose leadership would be open to adopting the intervention. Nearly all teachers within recruited schools participated in the study (in part due to the generous stipend provided by NWP for participation). Further, the schools are located in five different states and represent a mix of high-needs urban and rural schools. We therefore also estimate models using random teacher effects, which model the teacher effects as normally distributed using a Bayesian shrinkage estimator to "borrow" data from teachers with more precisely estimated effects.

We estimated all multilevel models using the Stata 14.2 *mixed* command. The models used restricted maximum likelihood estimation and the Kenward-Roger method to compute degrees of freedom for the models and calculate *p*-values to adjust for the relatively small sample size at the teacher and classroom levels (Kenward & Roger, 1997; Schaalje, McBride, & Fellingham, 2002).

Findings

This section describes our findings related to program implementation (both fidelity and treatment-control contrast), teacher-practice outcomes, and student-learning outcomes.

Fidelity of program implementation (FOI)

C3WP was implemented largely as intended. The implementation measures focused on professional development duration, content, and strategies. Nearly all sites met their thresholds for implementation fidelity over the course of the full school year.

- For the **duration and breadth** component, fidelity of implementation was defined as 80% of 7th and 8th-grade study teachers participating in 45 hours or more of professional development. Sites could meet this threshold if either 85%+ of teachers participated in 35

hours of professional development, or 75%+ of teachers participated in 40+ hours of professional development. All six sites met this threshold.

- The **content** component was composed of focus on argument and formative assessment components:

  Focus on argument was defined as 80% of 7th and 8th-grade study teachers participating in 36 hours or more of professional development focused on argument. Sites could meet this threshold if either 85%+ of teachers participated in 25 hours of professional development, or 75%+ of teachers participated in 30+ hours of professional development. All six sites met this threshold.

  Formative assessment required the introduction of the UST in professional development and the use of the UST online platform twice. Sites met this threshold if one professional development, even in summer or fall 2016, focused on the UST, and if analysis of student work was submitted at least twice prior to the administration of the outcome measure. All but one site met this threshold.

- The **classroom enactment** component was defined as at least 50% of professional development provided to 7th and 8th grade study teachers including classroom enactment of the C3WP tools and materials. Such enactment could include: (1) demo lesson, (2) designing tasks/assignments, (3) planning for classroom implementation, (4) analyzing student work, (5) analyzing student work, (6) Modeling instruction with teachers, (7) co-teaching/co-planning; and (8) debriefing classroom implementation. All sites met this threshold

Implementation fidelity thresholds were set at an annual level, which could, potentially, mean that sites only met fidelity after student posttest data were collected. We therefore also examined whether sites met implementation fidelity prior to the collection of outcome data. When examining the extent to which sites met these thresholds before the outcome measure was collected, we found that two sites did not meet the overall duration and breadth measure. The number of sites meeting the other FOI measures does not change.

We also examined the content on which sites chose to focus (reported to allow the research team to determine whether sessions focused on argument). All sites covered routine argument writing, the UST, and multiple instructional resources. While sites chose from multiple instructional resources, all sites introduced teachers to the resources that covered informal arguments and connecting evidence to claims. At the end of the year, two of the six sites introduced teachers to extended argument units intended to support students in a culminating argument.

Teacher practice outcomes

C3WP seeks to influence both the nature of teachers' instructional practices and the amount of time they spend teaching source-based argument writing. As described above, we administered a log that measured teachers' daily instructional practice during two weeks of the school year: once in fall and once in spring.

Teachers were substantially more likely to ask students to write during their C3WP classes than in their control classes (81% of days in C3WP classes v. 63% of days in control classes, $p < .001$) (Exhibit 4). When writing took place, students wrote for longer times in C3WP classes

(29 minutes v. 24 minutes, $p < .01$), although they were no more likely to write from source material (95% of days v. 93% of days, $p > .05$).

On days that teachers asked students to write in a class, the log provided teachers with a list of options for the purposes of writing meant to reflect genre choices or instructional purposes (e.g., argument, narrative, writing to learn) and were allowed to select all options that applied to that class's writing work. When asked to write, C3WP classes were much more likely to be asked to write argument (64% of days in C3WP classes v. 19% of days in control classes, $p < .001$) (Exhibit 5). Teachers were also more likely to select "to persuade others" for a purpose in their C3WP classes, although this happened infrequently and may indicate that some teachers see "argument" and "persuasive" as similar genres (3% v. 1%, $p < .05$). C3WP classes were less likely than control classes to write literary analysis (4% v. 11%, $p < .05$). C3WP classes were also somewhat less likely than control classes to write to learn (18% v. 24%, $p < .1$), write in preparation for standardized tests (6% v. 9%, $p < .1$), and write to summarize (4% v. 8%, $p < .1$), although these differences only approached statistical significance.

When writing argument, teachers nearly always expected students in both classes to use evidence (96% of days in C3WP classes v. 97% of days in control classes, $p > .05$) (Exhibit 6), and the evidence sources were almost always non-fiction (97% v. 96%, $p > .05$). However, teachers were nearly three times as likely to ask C3WP classes to draw from multiple sources relative to their control classes (76% v. 28%, $p < .001$).

When writing argument, teachers generally tended to ask students to practice similar skills in both C3WP and control classes, although C3WP students may have had greater expectations to revise particular elements of their arguments. Teachers were somewhat more likely to ask students in C3WP classes to focus on engaging in discussion about an argument topic (21% of days in C3WP classes v. 9% of days in control classes, $p > .1$) (Exhibit 7) and somewhat less likely to ask C3WP students to practice supporting a claim with evidence (72% v. 85%, $p < .1$), though these differences only approached statistical significance. While expectations that students revise their arguments were somewhat infrequent in both classes, students in C3WP classes were more likely to be expected to revise their claims or thesis statements (15% v. 4%, $p < .05$) and their selection of evidence (14% v. 5%, $p < .05$) relative to students in control classes. Students in both classes were equally likely to be asked to revise for commentary (18% v. 11%, $p > .05$) and for organization (12% v. 7%, $p > .05$).

Exhibit 4. Teacher Practice: Instructional Time Spent on Writing

	C3WP	Control		Log *N*	Class *N*	Teacher *N*
Did you ask students to write during this class?	81%	63%	***	859	58	29
If students wrote, did the task ask students to analyze, respond to, use text?	95%	93%		577	57	29
If students wrote, minutes student spent working on writing during class	29	24	**	576	57	29

NOTE: ~p < .10, *p < .05; **p < .01; ***p < .001; point estimates for models predicting binary outcomes were transformed into percentage points for ease of interpretation, and represent the predicted practice for the average class in the sample for which the question was asked. Teachers were not asked log questions if preceding questions indicated the answer was no (e.g., teachers were only asked about the facets of their writing practice on days that students wrote in that class).

Exhibit 5. Teacher Practice: Purpose of Writing Instruction (When Writing Took Place)

For what purpose(s) did this student write today? Please respond based on all the pieces of writing this student worked on. (Select all that apply)	C3WP	Control		Log *N*	Class *N*	Teacher *N*
To reflect on an experience or topic	11%	10%		577	57	29
To help the student learn	18%	24%	~	577	57	29
To practice conventions and usage	5%	4%		577	57	29
To express him- or herself creatively	3%	5%		577	57	29
To recount an event through narrative	2%	2%		577	57	29
To present information	15%	14%		577	57	29
To persuade others	3%	1%	*	577	57	29
To make an argument	64%	19%	***	577	57	29
To gain practice with forms of writing encountered on standardized tests	6%	9%	~	577	57	29
To gain practice with writing literary analyses	4%	11%	**	577	57	29
To help the student build awareness of his or her own growth and learning	3%	3%		577	57	29
To summarize	4%	8%	~	577	57	29
Other	1%	1%		577	57	29

NOTE: ~p < .10, *p < .05; **p < .01; ***p < .001; point estimates were transformed into percentage points for ease of interpretation, and teachers were only asked these questions if they answered that students wrote during that class. Answers therefore represent the predicted writing practice for the average class in the sample on a day that class wrote.

Exhibit 6. Teacher Practice: Use of Evidence and Source Material (When Argument Writing Took Place)

	C3WP	Control		Log *N*	Class *N*	Teacher *N*
Did the student's work on argument require them to use evidence?	96%	97%		254	48	29
Was the primary type of evidence the student was expected to use non-fiction?	97%	96%		245	48	29
Did the student use more than one source?	76%	28%	***	229	47	29

NOTE: ~p < .10, *p < .05; **p < .01; ***p < .001; point estimates were transformed into percentage points for ease of interpretation. Teachers were only asked these questions if they answered that students wrote during that class. Teachers were only asked about the type of evidence and number of sources students used if they answered that students used evidence. Answers therefore represent the predicted writing practice for the average class in the sample on a day that class wrote argument.

Exhibit 7. Teacher Practice: Argument Skills Practiced (When Argument Writing Took Place)

Did the student work on any aspects of argument during this class? (Select all that apply)	C3WP	Control		Log *N*	Class *N*	Teacher *N*
Engaging in discussion about an argument topic	21%	9%	~	254	48	29
Identifying an argument from daily life (e.g., ads, cartoons, op-eds)	4%	1%		254	48	29
Identifying a claim and/or evidence in text	26%	18%		254	48	29
Responding to a claim in a text	11%	7%		254	48	29
Revisiting a claim based on new reading or information	7%	2%		254	48	29
Generating ideas for future argument writing	5%	2%		254	48	29
Developing a claim	43%	49%		254	48	29
Evaluating the credibility of an argument	6%	3%		254	48	29
Supporting a claim with evidence	72%	85%	~	254	48	29
Elaborating upon evidence used to support a claim	48%	35%		254	48	29
Designing the organization of an argument	17%	15%		254	48	29
Revising the claim or thesis	15%	4%	*	254	48	29
Revising the selection of evidence	14%	5%	*	254	48	29
Revising for more effective commentary	18%	11%		254	48	29
Revising for organization	12%	7%		254	48	29

NOTE: ~p < .10, *p < .05; **p < .01; ***p < .001; point estimates were transformed into percentage points for ease of interpretation. Teachers were only asked these questions if they answered that students engaged in argument writing during that class. Answers therefore represent the predicted writing practice for the average class in the sample on a day that class wrote argument.

Student learning outcomes

Results of the impacts of a one-year implementation of C3WP on student achievement, as measured by their performance on the ACW-SBA are provided in Exhibit 8. For each component, all point estimates are positive, but both the size and significance of the estimation is sensitive to model specification. Nearly all estimations are either statistically significant ($p < .05$) or marginally significant ($p < .1$). We interpret these outcomes as indicative This suggests that although schools may see results from C3WP in a single school year, a longer-term investment may produce a greater impact.

Exhibit 8. Impacts on AWC-SBA

	(1)		(2)		(3)	
Content	0.13	~	0.14	~	0.17	*
	(0.06)		(0.08)		(0.08)	
Structure	0.13	*	0.13		0.15	~
	(0.06)		(0.08)		(0.08)	
Stance	0.12	~	0.15	~	0.17	*
	(0.06)		(0.07)		(0.07)	
Conventions	0.13	*	0.14	~	0.15	*
	(0.06)		(0.07)		(0.07)	
Block/ teacher N	30		30		30	
Classroom N	60		60		60	
Student N	939		939		939	
Teacher-by-class FE	x					
Block FE			x			
Block RE					x	
Fixed Treatment Effect			x		x	

~$p < .1$; *$p < .05$; **$p < .01$; ***$p < .001$

References

Arshan, N. L., Park, C. J., & Gallagher, H. A. (2018). Impacts on Students of a Short-Cycle Implementation of the National Writing Project's College, Career, and Community Writers Program. Menlo Park, CA: SRI International.

Bang, H. J. (2013). Reliability of National Writing Project's Analytic Writing Continuum Assessment System. *The Journal of Writing Assessment*, *6*(1).

Gallagher, H. A., Arshan, N., & Woodworth, K. R. (2017). Impact of the National Writing Project's College-Ready Writers Program in High-Need Rural Districts. *Journal of Research on Effectiveness in Education*, *10*(3), 570–595. DOI: 10.1080/19345747.2017.1300361

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. Biometrics, 53, 983–997.

Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. Journal of Agricultural, Biological and Environmental Statistics, 7(4), 512–524.