Johns Hopkins School of Education | Institute for Education Policy

Implementing Wit & Wisdom: An Evaluation
Research Report
June 2021

We compare the short-term literacy outcomes of students whose teachers began using the Wit & Wisdom curriculum at the beginning of the 2018-2019 school year, to students in nearby schools whose teachers continued using other curricular materials. Estimates from quasi-experimental propensity score matching show positive gains in 1st- through 3rd-grade students' Text Reading and Comprehension (TRC) *scores and in 4th- and 5th-grade students' ELA state test scores. The findings suggest that curricula can improve students learning, even in the first year of implementation.*

Research Overview

A child's success in school is dependent upon many factors. The most important in-school factor for students' outcomes is the teacher (Chetty et al., 2014; Rivkin et al., 2005b), and evidence shows that teachers can be more effective with more support (Jackson & Makarin, 2018). One form of support is the curricular materials teachers use to plan and deliver their lessons.

This paper studies the impact of switching to a new curriculum on students' literacy outcomes. Previous research shows that curricular materials impact teaching practices and students' learning in different ways (Chingos & Whitehurst, 2012; Polikoff & Koedel, 2017). In addition, curricular choices are a particularly important policy lever, because every school uses curriculum, many schools update or change their curriculum regularly, and there is no clear relationship between the cost and quality of curriculum. Therefore, choosing a high-quality curriculum could be a cost-neutral reform that positively impacts student outcomes.

However, despite the widespread use of curriculum, studying the impact of a specific curriculum on student learning is not straightforward. One reason for this difficulty is that many state agencies do not collect data on which curriculum schools use (Bhatt & Koedel, 2012), making it difficult for researchers to add "curricular quality" to the list of variables considered when assessing student learning. In addition, research shows that teachers' use of curricula is often inconsistent and haphazard (Opfer et al., 2016), making it difficult to determine whether the curriculum is ineffective or simply not used with fidelity. Curriculum quality is also multi-dimensional and complex. Websites such as EdReports and What Works Clearinghouse (WWC) have developed their own methods of defining curricular quality. These websites rate curricula very differently, however, and do not contain exhaustive lists.

In this paper, we estimate the impact of Wit & Wisdom (W&W) on student outcomes. We note that EdReports has evaluated W&W based on EdReports' criteria (i.e., text quality, building of knowledge, and usability), and determined W&W to meet its highest curricula quality rating. However, this paper adds to our understanding of W&W by studying the plausible impact of the curricula on student learning using WWC research guidelines.

The most rigorous design for evaluating the impact of a curriculum on student outcomes is through a randomized control trial (RCT). In this study design, some schools would be randomly selected to use a new curriculum, and other schools would not. However, few districts and schools are willing to allow researchers to randomly select their curriculum, making this design unfeasible.

Instead, this paper uses a quasi-experimental method called propensity score matching (Rosenbaum & Rubin, 1983), which approximates experimental effect sizes by creating an apples-to-apples comparison. Specifically, we match a student exposed to the intervention to a student not exposed to the intervention, but who is as identical to the intervention student as possible. This matching method produces *plausibly* causal estimates, meaning that results can likely be attributed to the intervention and not to other differences between the two groups under comparison.

In this paper we take advantage of the adoption of a high-quality ELA curriculum, W&W, in a single school district to examine the impact of a new curriculum within a district after the first year of implementation. Specifically, we use longitudinal administrative data and propensity score matching to estimate the difference in literacy outcomes for students in kindergarten through fifth grade. Our purpose is to provide evidence and information to those seeking knowledge about choosing a high-quality ELA curriculum, or W&W specifically.

We find that W&W students experience increased literacy-related scores in the first year of implementation, on average, with some variation by grade and test. While this evaluation was designed as a multi-year study, school disruptions from COVID-19 complicated data collection in the end of the second year. Therefore, while the study may continue into future years, this evaluation considers data from the first year of implementation only, thus providing evidence of short-term impacts.

Literature Review

High Quality Curriculum

While research on curricular efficacy is surprisingly scant, what does exist generally finds that implementing a high-quality, knowledge-rich curriculum can increase student learning and equity across classrooms (Bhatt et al., 2013; Bhatt & Koedel, 2012), while also providing support to teachers. Specifically, research shows that teachers affect students' outcomes more than any other in-school factor (Adnot et al., 2017; Chetty et al., 2014; Rivkin et al., 2005b), but does not offer clear answers about why some teachers are more effective than others (T. J. Kane et al., 2008; Rockoff et al., 2011; Tyler et al., 2010).

What we do know, however, is that implementing a high-quality, knowledge-rich curriculum that provides rigorous lessons and materials that are aligned to standards can improve both teacher quality and student outcomes (Dolfin et al., 2019; Hattie, 2009). High-quality curricula are correlated with significant increases in student learning and equity across classrooms. So, while teachers are the single most important in-school determinant to students' success, it is the curricular materials teachers use that shape what and how teachers teach (Chetty et al., 2014; T. Kane et al., 2016; Rivkin et al., 2005a). Further, a well-implemented curriculum can ensure more coherent instruction and better allocation of instructional time, and also encourage teacher collaboration (Dougherty, 2016; Jackson & Makarin, 2018).

For example, in one large-scale randomized controlled trial, students learned approximately 38 days more in math using one of three curricula, compared to a fourth curricula (Agodini et al., 2013). This particular study measures impacts on student learning after one year, but students benefit from a better curriculum year after year, thus potentially compounding benefits. In a study of the effectiveness of mathematics curricula, Bhatt and Koedel (2012) find that of two similar mathematics curricula, one is more effective, and that the difference in effectiveness appears to be due to design features of the curriculum. Similarly, in a study of mathematics curricula across 39 schools assigned one of four math curriculum, Agodoni et al. (2009) finds that two curricula produce statistically significantly higher math achievement for first-grade students. However, in 2019, leading researchers – again, considering mathematics textbooks - find no differences in fourth- and fifth-grade student math test scores (Blazar et al., 2019). Despite this discrepancy with findings from earlier studies, the researchers from this study call for continued research into the effectiveness of curriculum, because it has the potential to impact large populations of students quickly and efficiently (Blazar et al., 2019; Dolfin et al., 2019).

Looking at outcomes in ELA specifically, an experimental study (Borman et al., 2008) of elementary classrooms finds that classrooms randomly assigned to use a higher-quality curriculum had significantly higher reading composite, vocabulary, and reading comprehension scores. A study of the effectiveness of a standards-aligned ELA curriculum, paired with teacher learning supports, also finds positive impacts on teacher practices such as engaging students in discussions, writing about their learning, and encouraging higher-order thinking. In addition to evidence of improved teacher practice, the study found evidence of improved student outcomes in the second year their teacher was exposed to the curriculum and supports (Dolfin et al., 2019).

Equity

Using high-quality curriculum also has important implications for equity, by ensuring that all students have access to rigorous lessons. Recent research from The New Teacher Project (TNTP, 2018) shows that low-income students in the United States are half as likely to receive lessons with grade-level content than their more advantaged peers. The same report also finds that students learn more when they are given high-quality, grade-level lessons. For example:

> In classrooms where students had greater access to grade-appropriate assignments, they gained nearly two months of additional learning compared to their peers… The relationships between the resources and student outcomes were even stronger in classrooms where students started the year off behind. When students who started the year behind grade level had access to stronger instruction, for example, they closed gaps with their peers by six months; in classrooms with more grade-appropriate assignments, those gaps closed by more than seven months (Introduction, Opportunity Myth).

Remedying inequities in access to rigorous materials, then, is a clear path to greater educational equity.

In addition to the benefits for all students within a school, adopting the same high-quality curriculum throughout the district can be particularly beneficial for low-income students. Specifically, low-income students are significantly more likely to change schools than their more privileged peers, a difference that transcends race (Parke & Kanyongo, 2012). Therefore, when an entire district adopts the same high-quality curriculum, higher-mobility students benefit not only from the better materials, and grade-level lessons, but also from the continuation of systemic instruction between schools.

Teacher Efficacy

Better materials also help teachers. High-quality instructional materials have been shown to provide more support to teachers; in particular, they have been found to strengthen the performance of struggling teachers (Jackson & Makarin, 2018). For example, a recent randomized control trial (Jackson & Makarin, 2018) compares student test scores amongst teachers who: (1) were given access to high-quality materials; (2) were given access not only to the high-quality materials, but also to support in using the materials; or (3) received neither materials nor support. When teachers had access to high-quality curriculum only, students increased their learning by approximately 35 days. However, when teachers were also given support in the use of the high-quality materials, students increased learning by roughly 55 additional days in comparison to classrooms with no additional materials or support.

This same research shows that high-quality curriculum is especially supportive for struggling teachers. Students with the weakest teachers showed the most gains: students in classrooms with teachers ranked in the bottom 60% gained 64 additional days of learning, on average, when their teachers were given access to both high-quality materials and professional development/coaching support.

On balance, research thus provides support for the view that switching to high-quality curricula and using them with fidelity is a cost-effective reform that benefits both teachers and students. Unfortunately, most teachers in the United States spend hours each week selecting and creating instructional materials – even if they are provided with a preferred district curriculum. This patchwork curriculum-building creates multiple problems. There is no consistency among what is taught within a district, within in a school, or even within the same grades in the same school. Further, the time used to create materials could have been used for preparing for classroom instruction. High-quality curriculum increases teacher efficiency by decreasing the burden of planning. The RAND Corporation's American Teacher Panel, a nationally representative sample of 1,705 teachers, shows that teachers spend an average of four or more hours planning per week (Opfer et al., 2016). According to teachers, searching for curricular materials accounts for most of this time. High-quality curriculum, therefore, allows teachers to switch from planning what to teach to how best to teach.

Challenges with Fidelity and Time

Given all these benefits one might wonder why every district does not immediately adopt high-quality materials. There are, of course, a variety of real-world challenges that districts and schools face when it comes to selecting, purchasing, implementing, and supporting a new curriculum. Challenges can include a district purchasing but not supporting the use of the new curriculum through encouraging adequate building level buy-in, and not supporting teachers with aligned professional development (Wei et al., 2009). Teachers may not embrace the new curriculum, thus either not using it at all or not using it with the needed fidelity. A second set of issues has also emerged in practice: The full effect of a new curriculum may take time to realize (Nichols-Barrer & Haimson, 2013), that changing curriculum may even create temporary dips in student outcomes (LIFT & TNTP, 2019), and that schools may panic and halt implementation before its full potential has been reached.

Both qualitative case studies and quantitative research show that successfully implementing a new curriculum takes time. A 2019 report from Leading Innovation for Tennessee Education (LIFT & TNTP, 2019) which details the districts' process of implementing high-quality instructional materials, shows that at the beginning of implementation, teachers aligned just 4% of their classroom practices to Tennessee's ELA standards. After three years of implementation, roughly half of the classrooms showed partial or full alignment to state standards. These findings highlight the implementation challenge that inducing, training, and supporting teachers' effective use of standards-aligned curriculum takes time.

Quantitative research provides additional evidence of how long complete implementation can take: A three-year study of implementing a high-quality curriculum combined with professional learning (Nichols-Barrer & Haimson, 2013) showed small gains in the first year of implementation in reading scores (i.e., a 0.06 standard deviation increase) and no significant increase in math (i.e., a -0.02 standard deviation decrease). However, by the end of the three-year study, students who had experienced all three years of the curriculum showed consistently strong learning gains (i.e., a 0.16 standard deviation increase in reading scores and a 0.29 standard deviation increase in math scores). The cumulative impacts of this curriculum were equivalent to moving a student from the 50th percentile to the 56th percentile in reading, and from the 50th to 60th percentile in math after three years. Therefore, research shows that a single school year is typically not enough time to evaluate the full impact of curricular changes.

Evaluation Context

In fall 2018, a large suburban school district in North Carolina made the first step towards implementing a high-quality curriculum by adopting W&W in most of its elementary schools in 2018-19 academic year. Our evaluation follows the implementation of W&W in this district and leverages both qualitative and quantitative analysis to better understand the impact of the new curriculum on student literacy outcomes.

The Wit & Wisdom Curriculum

W&W is a Kindergarten through eighth grade curriculum published by Great Minds. Developed by teachers, it combines knowledge-building lessons and a framework for reading complex texts, combined with authentic art and literature. W&W also offers an integrated approach to literacy, in which students receive explicit instruction on reading, speaking, listening, and writing within the same module (i.e., unit) topic.

W&W is one of EdReports' highest-ranked ELA curricula. In the grades reviewed by EdReports, third through eighth grades, W&W received scores of 41 out of 42 for text quality, 32 out of 32 for building knowledge, and 34 out of 34 for usability in elementary grades, with similarly high ratings for middle grades.

School District Context

The district under study is located in North Carolina and encompasses a midsized city and the surrounding suburban and rural area (NCES, 2006). It is one of the larger districts in North Carolina (out of 115) and serves more than 50,000 students. The district also has a diverse student body, of which almost half are African American; roughly a quarter, white; approximately 15%, Hispanic; and more than three quarters qualify for free or reduced-price meals.

In fall 2018, the district adopted W&W in general education classrooms in 80% of its elementary schools. All struggling schools were required to adopt W&W; the remaining schools were offered the choice to adopt the new curriculum. While most schools adopted W&W at all grade levels, some chose to only adopt at a few grade levels. Therefore, for the purposes of this study, curricular adoption took place at the grade-level.

The partial adoption within the district, and no adoption in nearby districts, allows us to compare the ELA outcomes for students according to whether they attended a W&W adopting school and grade. The schools that did not adopt W&W and the school districts that geographically border the W&W adopting district serve as "comparisons" to the W&W "intervention" students.

There are more than 100 elementary schools in the *matched* sample—the combined sample of students in schools adopting W&W and the students in nearby comparison schools. The schools in the matched sample are located in a range of locales, from remote rural to a midsized city. Specifically, about 40% of students in the matched sample attend schools in a rural area; 40% from the outskirts of an urban area (i.e., towns or suburbs); and 20% from an urban area.

Figure 1, below, provides a description of students' demographic characteristics in the entire matched sample, as well as students in the intervention (i.e., students in W&W schools) and comparison schools (i.e., students in non-W&W schools).

Figure 1

Student Characteristics of Matched Sample

	Total	W&W	Non-W&W
Female (%)	0.49	0.49	0.49
African American (%)	0.33	0.49	0.24
White (%)	0.32	0.24	0.37
Hispanic (%)	0.19	0.14	0.21
Other (%)	0.16	0.13	0.18
ELL (%)	0.08	0.04	0.10
Disabilities (%)	0.13	0.14	0.13
Qualify for FRL (%)	0.78	0.84	0.75
Number of Students	36,904	13,116	23,788

Figure 1 shows that the intervention and comparison matched samples are not the same for every student characteristic. While both samples have the same ratio of females, and similar percentages of students with disabilities, the W&W schools have a higher percentage of African-American students and students who qualify for free-and-reduced lunch, but lower percentages of white, Hispanic, and students learning English. However, students do have similar average baseline test scores, as required for research reporting standards by What Works Clearinghouse (WWC), and discussed in greater detail, below. In addition, using propensity scores to estimate treatment effects is equivalent to controlling for these differences in observed characteristics (Powell et al., 2019) .

Data

In this paper, we combine multiple sources of quantitative, and two sources of qualitative, data. Specifically, we use administrative data collected from the district, state, or federal government for our quantitative sources, and we collected qualitative data in order to better understand the districts' context and experiences.

First, we use school-, teacher- and student-level longitudinal administrative data reported to the state from the 2017-2018 and 2018-2019 school years. The student-level information includes students' school and teacher assignment and demographic characteristics (e.g., race, age, and if the student has a disability). These data also provide several outcome measures of student learning. This paper uses two sources of student test scores due to variance in available scores by grade-bands. First, we focus on one measure of early literacy for first- through third-grade students—the Text Reading and Comprehension (TRC) Benchmark Measure portion of the mCLASS: Reading 3D assessment. Second, we use the North Carolina End-of-Grade Reading test scores for fourth and fifth-grade students.

The TRC is a leveled reading assessment administered by the classroom teacher three times a year and is used to determine a student's reading level, which combines both decoding and reading comprehension. Reading levels are reported in the following way: after two pre-reading levels (i.e., Print Concepts and Reading Behaviors), students receive a rating from A-Z and a proficiency designation of "instructional" or "independent" within that alphabetical level. The proficiency levels indicate either that students need instructional support to fully comprehend and accurately read a book at an alphabetical reading level (i.e., "instructional"), or that students can accurately read and fully comprehend the book independently (i.e., "independent"). Therefore, a student reading at an instructional level (e.g., instructional G) is not as strong a reader as one reading at the same level independently (e.g., independent G). Also, note that when students can independently read on a given level (e.g., G), they are automatically tested on the next alphabetical level (e.g., H). Therefore, an independent reading level means that the student can independently read at that level (e.g., independent G), but is not yet ready for the next level (e.g., instructional H). As such, a 1-point score increase in this TRC measure indicates either a move from instructional to independent within the same alphabetical reading level, or a move from independent in one alphabetical reading level to instructional in the next level.

Second, we use the North Carolina End-of-Grade Reading test, a multiple-choice test aligned to North Carolina state standards. In the elementary grades, the test is administered in third through fifth grades and requires students to read selections of text and answer related questions. All tests are administered within the last 10 days of the school year, providing a measure of students' year-end reading performance.

In addition to student-level data, we incorporate administrative data about schools and teachers. This information includes teachers' responses to questions about their school-level work conditions from the 2018 Teacher Working Conditions Survey, a survey administered statewide every two years. The survey provides measures such as use of time, teacher leadership, professional development, and instructional practices and support. This administrative data also includes detailed teacher-level information about payroll information, teacher experience, and education.

Finally, we add publicly available data from the National Center for Education Statistics (NCES) to incorporate school-level characteristics within the W&W-adopting district and its bordering districts. These data include information such as school size, locale, and the percentage of students who qualify for free-and-reduced lunch within each school.

In addition to the quantitative data sources, we add two qualitative data sources, to help us better understand and interpret our results. Specifically, we collected survey response data from all district teachers, as well as from school leaders. These survey data provide evidence of curriculum use in schools, fidelity of use of the W&W curriculum in schools where it was adopted, information on teacher practices, and information on professional development. These data also provide us with an insider look at what teachers did and thought, and what sorts of supports they received during the year they implemented W&W.

We also observed district teachers in classrooms in the spring of the first year of implementation. Specifically, we conducted observations in 88 classrooms in April and May of 2019 in 15 randomly-selected schools (12 that used W&W and three that did not). The observers paid close attention to how teaching and learning in classrooms aligned to the W&W and standards-aligned instructional practices using an observation rubric. The rubric combined elements from W&W's Implementation Expectations; The North Carolina Department of Public Instruction's Instructional Practices for ELA Practitioner Checklist; and the Instructional Practices Guide (IPG), a college- and career-readiness, standards-aligned observational rubric created by Student Achievement Partners. The survey and observation data provide context and rich evidence as to the implementation fidelity of the W&W curriculum within schools and across the district.

Frameworks for Understanding Research

The Every Student Succeeds Act (ESSA)

ESSA offers guidance to educational agencies for selecting interventions supported by strong evidence (USDOE, 2016). While ESSA does not mandate adherence to this guidance, some competitive grant programs do require the use of "evidence-based" interventions (i.e., those supported by higher levels of evidence). ESSA outlines four "tiers" of evidence. Tier one provides the strongest evidence and requires significant positive outcomes from at least one well-designed and implemented randomized control trial, and no studies showing significant negative outcomes. The second tier of evidence is moderate and requires significant positive outcomes from a well-designed and implemented quasi-experimental study, and no studies showing significant negative outcomes. The third tier of evidence is promising, and requires positive significant outcomes generated by a correlational study, and no studies showing significant negative outcomes. The fourth and final level of evidence must demonstrate a rationale that the intervention is likely to improve student outcomes.

What Works Clearinghouse (WWC)

For additional support in selecting evidence-backed educational interventions, the WWC serves as a quality-control monitor for the level of scientific evidence on education programs and products. WWC also provides guidance and guard rails to researchers through additional technical requirements, requiring that research meets the most rigorous methods possible, and thereby provides a quality control mechanism for the evidence the study generates. WWC also reviews research according to these rigorous standards, thus attempting to separate well-designed and well-implemented studies that are more likely to have dependable findings, from those that are less reliable. They then present the outcomes to education stakeholders looking for reliable evidence. While ESSA does not require that an intervention be approved by WWC, WWC standards certainly meet those of ESSA.

Methodology

This paper uses quasi-experimental statistical techniques (i.e., ESSA Tier II evidence), and follows standards defined by WWC, as well as other commonly established methods. The logic behind our methodology is as follows: In order to estimate the impact of an intervention on student outcomes, we ideally want to compare the outcomes for the student exposed to W&W to what would have happened to that same student without the intervention. This comparison is clearly impossible. Therefore, for each student exposed to the W&W, we look for a student not exposed to W&W who looks as similar as possible, before the intervention started.

This method is called "matching," and produces plausibly causal estimates. That is, when all relevant differences between the W&W and comparison groups that affect literacy outcomes are captured in the model, then the model estimates the impact of W&W. However, the model cannot account for differences that are unobserved in the data and impact literacy outcomes.

We can formally define the effect of the treatment (i.e., W&W) as:

$$\tau_1 = E(y_1 - y_0) \tag{1}$$

Equation 1 states that our goal is to estimate the average treatment effect (τ_1), which is the expected difference in outcomes between receiving the treatment (y_1) and what the outcome would have been without the treatment (y_0).

The challenge with estimating the average treatment effect, however, is that for each individual student, i, we do not observe both outcomes. Rather, we only ever observe realized outcome: (y_{1i}) if a student received the intervention or (y_{0i}) if they did not. Therefore, we need to estimate what a students' outcomes might reasonably have been, if they had been in the other group. Our matching method finds observationally similar students to do this.

One of the most critical methodological aspects of matching is selecting which characteristics students are matched on. One could conceive, for example, of matching students with the same test scores before the introduction of W&W in the classroom. Using this single student characteristic, test scores, would likely result in many potential matches for each intervention student. However, just because students have the same test score, does not mean that they are otherwise very much alike. On the other hand, matching students with similar family backgrounds, similar schooling experiences, and similar academic levels before the intervention might produce a very good match, but it is unlikely such a match exists for most students.

Propensity scores balance the two concerns of matching on enough characteristics and of finding matches that look similar in the data, by estimating a single number from the matching characteristics. That single number, or propensity score, is the estimated probability that an individual will receive the treatment, given their characteristics. Therefore, propensity scores allow for the use of many characteristics, and reframes what "similar" means.

While propensity scores allow for the inclusion of as many characteristics to match upon as the data allow, there are guidelines for selecting which characteristics to (Holmes, 2014). Specifically, the characteristics used to match students should be related to both selection into treatment and the outcome of interest, since these are the confounding characteristics that we need to account for. However, we must also be careful that any characteristic included in the model are not affected by the selection process (i.e., endogenous).

The final endogeneity requirement is an important methodological requirement for meeting WWC guidelines. This model meets these WWC requirements, because many measures were collected before the beginning of the intervention (e.g., the work conditions survey and state tests were given the spring before the intervention), and all other information was measured at the beginning of the intervention (e.g., school sizes, teacher's salaries, and student-teacher ratio are measured at the beginning of the school year), thus making them unlikely to have changed due to the selection process. The WWC Standards Handbook clearly states that under these conditions, covariates are not potentially endogenous (What Works Clearinghouse, 2020, p. 33).

Our rich data set allows us to consider hundreds of potential measures at the district, school, and classroom levels that might impact selection and student outcomes. In order to first look at the relationship between characteristics and selection into treatment, we employed a machine learning technique called Lasso regression (least absolute shrinkage and selection operator) (Tibshirani, 1996). Lasso regression can be used to shrink the coefficients of variables that are not predictive of an outcome and can therefore be used in model selection. In this case, we used Lasso regression to select which pre-intervention measures were most predictive of selection into the W&W intervention.

Next, we used prior research to provide evidence as to which of the characteristics we found using Lasso (i.e., predictive of the intervention), were also predictive to student learning (i.e., our outcome of interest). We found evidence that a school's locale (e.g., rural, urban, etc.) (Jimenez-Castellanos, 2010; Lafortune et al., 2018; Lane et al., 2018; Shores & Ejdemyr, 2017); size (Egalite, 2017) ; student-teacher ratios (Bosworth, 2014; Li & Konstantopoulos, 2017); percentage of students eligible for free or reduced lunch(Breger, 2017; Carnoy & García, 2017; Reardon, 2011); and support for collaborative planning (Reeves et al., 2017) are all predictive of differences in student outcomes. In addition, we found a research basis linking teachers' pay (Hanushek et al., 2019; Hendricks, 2014; Pham et al., 2021), experience, and education to student outcomes. We found a weaker research basis linking teachers of foreign decent (Seah, 2018) ; support for new teachers (Ronfeldt & McQueen, 2017); and teachers' desire for PD aimed at helping students with disabilities to student outcomes (Feng & Sass, 2013; Fischer et al., 2018). However, there is some research suggesting that immigrant teachers systematically have different impacts on students, and that PD aimed at helping students with disabilities can be beneficial for general education teachers, for example. Therefore, we also included these characteristics in our model.

However, some characteristics that were predictive of the intervention had little to no evidence that they were linked to student outcomes. For example, we found that teachers' responses to questions about efforts to reduce routine paperwork; the amount of time teachers spend on school related activities outside of school hours; and responses regarding whether the school board was elected, were all predictive of the intervention, but not linked to student outcomes in prior literature. Researchers have shown that including characteristics that are correlated with the intervention selection, but not the outcome, do not add precision to the estimates and can lead to dropped observations (Austin et al., 2007). Therefore, we do not include any of these characteristics in our propensity score model.

Thus, we used a combination of empirical methods and prior research in order to identify the characteristics, or variables, that form the basis of our propensity score model. We define the estimated propensity score of student i, $\hat{\mathrm{p}}(x_i)$, as the conditional probability of receiving the intervention (i.e., taught using W&W), t, given the student's observed characteristics ($\boldsymbol{x}_i$) as:

$$\hat{\mathrm{p}}(x_i) = \Pr(t_i = 1|\boldsymbol{x}_i) \qquad (2)$$

As noted above, the vector of observed student characteristics, $\boldsymbol{x}_i$, includes a variety of students' school-level characteristics, including the number of students the school serves; the amount of support first-year teachers receives; the amount of collaborative planning time at the school; teachers' perceived need for PD for students with disabilities; the school's locale; the percentage of students qualifying for free-and-reduced lunch in the school; and student-teacher ratios. Observed characteristics also include students' teacher characteristics, including education level, salary, teaching experience, and if the teacher participates in a foreign teacher program such as the J-1 Visa Teachers Program. The model also includes student characteristics that are important predictors of student outcomes, including grade, gender, race, pre-intervention test scores, disability status, ELL status, and age. We then combine these characteristics to estimate a propensity scores for each student using logistic regressions, which is the most common estimation method for estimate propensity scores (Powell et al., 2019).

Next, we use the nearest neighbor method with replacement to predict what a students' potential outcome would plausibly have been in the other group (e.g., the non-intervention group, for students in the W&W intervention group). Specifically, we calculate the average outcome of individuals with the closest propensity scores, but who received the other intervention condition. Formally, we predict the potential outcome ($\hat{y}_{ti}$) for student i as:

$$\hat{y}_{ti} = \begin{cases} y_i & \text{if } \ t_i = t \\ \dfrac{\sum_{j\in m} w_j y_j}{\sum_{j\in m} w_j} & \text{otherwise} \end{cases} \qquad (3)$$

Where y_i is student i's realized outcome and y_j is the outcome of student i's nearest neighbor. The least number of nearest-neighbor matches allowed are denoted as m, which we set to one, and w_j are frequency weights for student i's nearest neighbor (i.e., student j). Equation 3 essentially says that when we want the estimate of the intervention status the student experienced, we use the student's observed outcome. However, when we need an estimate for the intervention status the student did not experience, then we calculated the weighted average outcome from the at least m nearest neighbors who did experience the other treatment status.

Now that we have a plausible estimate for what the students' outcomes would have been if they had received the opposite treatment, we can estimate our average treatment effect ($\hat{\tau}_1$) as:

$$\hat{\tau}_1 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_{1i} - \hat{y}_{0i})}{\sum_{i=1}^{n} w_i} \tag{4}$$

Where w_i is the frequency weight of student i, and $\hat{y}_{1i}$ and $\hat{y}_{0i}$ are defined in equation 3. This equation essentially says that we estimate the average treatment effect by taking the weighted average of the difference between the observed and estimated outcomes for each individual.

In addition to the careful selection of characteristics into our model, WWC guidelines require that students in the intervention and comparison groups are similar by checking baseline equivalency between the two groups. Specifically, WWC requires that baseline equivalency is met for students' pre-test scores, only. Note that the pre-test score for 1st through 3rd grade students are TRC scores and the 4th and 5th grade scores are state test scores. Figure 2 shows the standardized mean difference calculations[1] for pre-test scores to establish baseline equivalency. WWC considers intervention and comparison groups as "equivalent" at baseline if standardized mean differences are 0.05 or less in absolute value. Mean differences at the grade levels presented in this paper meet this requirement.

Figure 2

Baseline Equivalency of Pre-Test Measures

	Baseline Equivalency	Sample Means	
	Standardized Mean Difference	W&W	Non-W&W
First Grade	-0.04	8.54	8.73
Second Grade	-0.02	19.55	19.70
Third Grade	-0.02	27.18	27.37
Fourth Grade	0.00	45.85	45.97
Fifth Grade	0.02	44.46	43.92

[1]These are calculated as the difference in means between the two sample groups, divided by the pooled standard deviation.

WWC guidelines also require adjustments to statistical significance when an intervention is adopted at a different level than reported estimates. In this study, the W&W intervention was adopted at the grade level (i.e., while most schools adopted the curriculum for all grades served in the school, some schools in the sample adopted only for some grade levels), and our analysis is presented at the student level. Therefore, we report statistical significance using the adjustments outlined in WWC's Procedures Handbook (What Works Clearinghouse, 2020, p. F-2). However, because WWC guidelines are not required to establish evidence for ESSA, and because the interpretation of the results is different using other rigorous methods, we also report our pre-WWC-adjusted statistically significant findings. These estimates follow an estimation procedure proposed by (Abadie & Imbens, 2016) and account for the fact that propensity scores are estimated from the data, as opposed to collected observable characteristics.

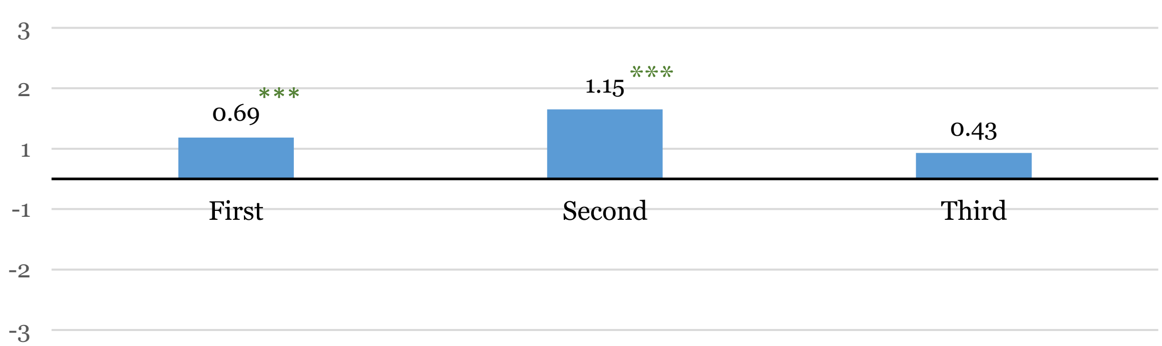
Results

The results in this section provide an estimate of the change in students' literacy scores plausibly caused by their teachers' switch to W&W. Figure 3 (below) shows the estimated intervention effect of W&W on students' early literacy skills, as measured by TRC scores, in comparison to matched students. Estimates from each grade are presented separately.

Figure 3

Estimated Impact of Wit & Wisdom Implementation on TRC Scores

Panel A: All W&W Teachers



NOTES—1. The statistical significance refers to the difference in the average student achievement between the Wit & Wisdom students in the study, and comparison students, using procedures defined by Abadie & Imbens (2016): ~p<.10, *p<.05, **p<.01, ***p<.001 and using WWC-proscribed adjustments: ~p<.10, *p<.05, **p<.01, ***p<.001

Figure 3 shows that the W&W implementation across all classrooms is associated with positive effects on students' early literacy, as measured by TRC scores in first- through third-grade classrooms. These estimates are statistically significant in first and second grades, when we estimate statistical significance using Abadie and Imbens (2016) estimators, but not so when we make additional adjustments required by WWC. Note that the magnitude of the estimated effect of implementing W&W are the same for both methods, but the statistical significance and interpretation of these effects are different.

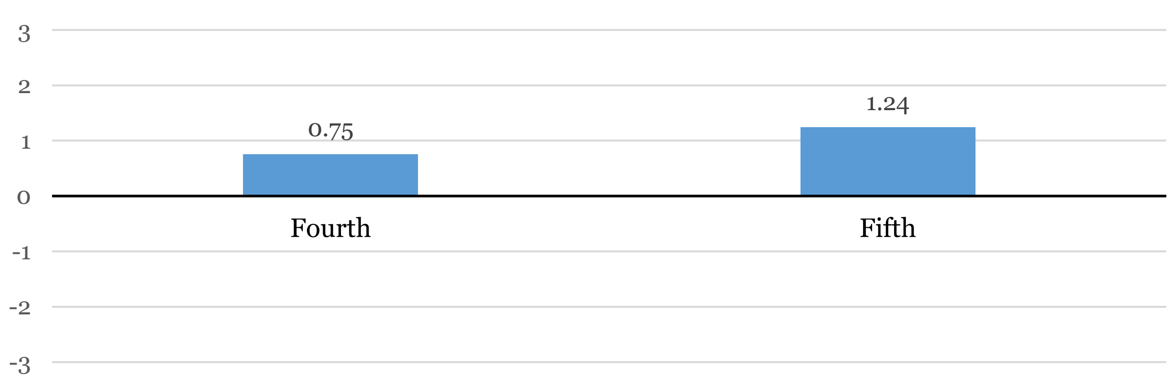
Estimates are interpreted as different than zero when they are statistically significant—that is, the difference between the two groups is unlikely due to chance. However, in the case of the adjusted estimates, the estimates are no longer considered statistically significant. This means we cannot rule out that these differences are due to chance.

Issues of how to interpret the significance levels of the estimates aside, the results show that all students experienced better TRC scores than they otherwise would have, after their teachers switched to the W&W curriculum. For example, first-grade students whose teachers started using W&W read at almost a one-point higher level (i.e., 0.69 point) than they would have if their teacher had not used W&W. This means that students in W&W classrooms moved approximately an additional level—from instructional to independent within the same alphabetical reading level (i.e., instructional J to independent J), for example, or from an independent reader to an instructional reader on the next alphabetical level (i.e., independent J to instructional K). Similarly, second-grade students experienced slightly more than a one-point increase (i.e., 1.15 point), on average, after their teachers switched to W&W. Finally, Figure 3 shows that third-grade students experienced almost a half-point (0.43) increase in TRC scores, after the switch to W&W.

Figure 4 presents the estimated impacts of switching to W&W on fourth- and fifth-grade students' state test scores. These estimates show that both fourth- and fifth-grade students experienced increases in their scores, but that these differences were not statistically significantly different from zero for either grade (or when using either method for estimating statistical significance). Specifically, in fourth grade, students from W&W classrooms experienced almost a one percentage point (i.e., 0.75) increase in state test scores when compared to their non-intervention peers. Similarly, fifth-grade students from W&W classrooms experienced slightly more than a one-percentage point increase (i.e., 1.24), than matched non-W&W students who were similar on observable characteristics.

Figure 4

Estimated Impact of Wit & Wisdom Implementation on Reading State Test Scores (Percentage Points)



NOTES—1. The statistical significance refers to the difference in the average student achievement between the Wit & Wisdom students in the study, and comparison students, using original Abadie & Imbens (2016) estimates: ~p<.10, *p<.05, **p<.01, ***p<.001 and using WWC adjustments: ~p<.10, *p<.05, **p<.01, ***p<.001

In order to better understand the estimates in Figure 4, we standardized our estimates. These calculations show that in fourth grade, W&W students gained an average of 0.03 standard deviations, as compared to their matched peers. In fifth grade, W&W students gained an average of 0.04 standard deviations, in comparison to their matched peers. These learning gains translate into roughly 11.5 additional days of learning in fourth grade and 19 additional days of learning in fifth grade, as research shows that fourth- and fifth-grade students literacy scores increase approximately 0.40 standard deviations on nationally normed tests (Hill et al., 2008).

Discussion

Recall that we also have extensive data gathered from stakeholder surveys and classroom observations during the first year of the W&W implementation. In order to better understand our findings, we explore these descriptive data.

The modest student gains found above comport with what we saw in classroom observations and with teacher survey responses. Specifically, both qualitative sources of evidence suggest that teachers used W&W, but struggled to teach some aspects of the curriculum. However, both teachers and observers noted promising changes in instruction and student learning.

Classroom observations and teacher survey responses suggest that almost all teachers in schools that adopted W&W used the curriculum. Specifically, 95% of teachers in schools that adopted W&W ("W&W schools"), indicated that they used W&W materials to plan their ELA lessons. In addition, 86% of teachers in W&W schools reported that they used W&W as their primary curriculum. School leaders' expectations for W&W use mirrored teachers' reported use. Specifically, all school leaders in W&W schools stated that they expect their teachers to use W&W and 90% expected their teachers to "always" use W&W.

However, both survey responses and classroom observations suggest that many teachers struggled to use W&W with fidelity. Only 40% of teachers in the W&W schools reported that they used every component of W&W "always." A larger percentage of teachers, 67%, used all W&W materials at least "most of the time." Despite encouraging reports of fidelity, classroom observers saw virtually no use of concluding lesson elements (i.e., land, wrap, or deep dives) and very few craft writing lessons, two key components of the W&W curriculum. Observers also noted that, given the lessons observed in the late spring, few teachers were likely to finish all four W&W modules before the end of the academic year. These findings suggest that both using materials with fidelity and pacing were challenges to teachers during their first year of W&W implementation.

Nonetheless, both observations and survey responses suggested promising changes in student learning. For example, the survey responses of teachers who used W&W as their primary curriculum stated that, since the adoption of W&W, their students wrote with more evidence from texts; increased their use of vocabulary; and had more knowledge about the world. In addition, teachers who used W&W more frequently reported observing more evidence of student learning in these dimensions, compared with teachers who used W&W less frequently.

Classroom observers noted differences among both teachers and students in W&W classrooms, compared to non-W&W classrooms. For example, teachers who used W&W required their students to provide precise answers with evidence from texts more consistently than teachers in classrooms not using W&W. In addition, students in W&W classrooms were more likely to read and re-read complex texts; practice speaking and listening skills through discussions; and practice writing. Students in W&W classrooms also used more sophisticated vocabulary and engaged in thoughtful discussions around challenging content.

Given these additional findings and support from the literature, it seems likely that improved fidelity and additional time for complete implementation will improve outcomes.

Conclusions

The results presented here were obtained with the most rigorous evaluation methods feasible, thus meeting ESSA's requirements for Tier 2, or moderate, evidence. Results suggest that the implementation of a high-quality ELA curriculum, W&W, in its first year had a positive effect on students' early literacy outcomes and a positive effect on upper-elementary students' state reading test scores, on average. Estimates in first and second grades are statistically significant when using standard estimation methods, but not under the adjustments required by WWC. Estimates in third through fifth grades are positive, but statistically insignificant. However, recall that meeting WWC standards is not a requirement of ESSA; therefore, the alternate estimates are also considered valid.

These technical differences aside, our estimates suggest that the implementation of W&W is associated with increased student literacy outcomes across all grades and scores evaluated, even in the first year of implementation, with students gaining a level on their TRC scores on average or increasing their ELA test scores by a percentage point. These are promising findings given prior research on both the challenges of implementing a new curriculum, and the length of time it often takes to fully realize the positive learning gains derived from implementing high-quality curricular materials.

Citations

Abadie, A., & Imbens, G. W. (2016). Matching on the Estimated Propensity Score. *Econometrica*, *84*(2), 781–807. https://doi.org/10.3982/ECTA11293

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis*, *39*(1), 54–76.

Agodini, R., Harris, B., Atkins-Burnett, S., Heaviside, S., Novak, T., & Murphy, R. (2009). *Achievement Effects of Four Early Elementary School Math Curricula: Findings from First Graders in 39 Schools.* (No. 2009–4052). National Center for Education Evaluation and Regional Assistance.

Agodini, R., Harris, B., Remillard, J., & Thomas, M. (2013). After Two Years, Three Elementary Math Curricula Outperform a Fourth. NCEE Technical Appendix. NCEE 2013-4019. *National Center for Education Evaluation and Regional Assistance*.

Austin, P. C., Grootendorst, P., Normand, S.-L. T., & Anderson, G. M. (2007). Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study. *Statistics in Medicine*, *26*(4), 754–768. https://doi.org/10.1002/sim.2618

Bhatt, R., & Koedel, C. (2012). Large-Scale Evaluations of Curricular Effectiveness: The Case of Elementary Mathematics in Indiana. *Educational Evaluation and Policy Analysis*, *34*(4), 391–412.

Bhatt, R., Koedel, C., & Lehmann, D. (2013). Is curriculum quality uniform? Evidence from Florida. *Economics of Education Review*, *34*, 107–121. https://doi.org/10.1016/j.econedurev.2013.01.014

Blazar, D., Kane, T., Staiger, D., Goldhaber, D., Hitch, R., Kurlaender, M., Heller, B., Polikoff, M., Carrell, S., Harris, D., & Holden, K. L. (2019). *Learning by the Book*. Harvard University. https://cepr.harvard.edu/files/cepr/files/cepr-curriculum-report_learning-by-the-book.pdf

Borman, G. D., Dowling, N. M., & Schneck, C. (2008). A multisite cluster randomized field trial of Open Court Reading. *Educational Evaluation and Policy Analysis*, *30*(4), 389–407.

Bosworth, R. (2014). Class size, class composition, and the distribution of student achievement. *Education Economics*, *22*(2), 141–165. https://doi.org/10.1080/09645292.2011.568698

Breger, L. (2017). Poverty and Student Achievement in Chicago Public Schools. *The American Economist*, *62*(2), 206–216. https://doi.org/10.1177/0569434516672759

Carnoy, M., & García, E. (2017). *Five key trends in U.S. student performance* (p. 62). Economic Policy Institute.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *The American Economic Review*, *104*(9), 2633–2679.

Chingos, M., & Whitehurst, G. (2012). *Choosing Blindly: Instructional Materials, Teacher Effectiveness, and the Common Core*. Brookings Institution. https://www.brookings.edu/research/choosing-blindly-instructional-materials-teacher-effectiveness-and-the-common-core/

Dolfin, S., Richman, S., Choi, J., Streke, A., Cheryl, D., Demers, A., & Poznyak, D. (2019). *Evaluation of the Teacher Potential Project*. Mathematica Policy Research. https://www.mathematica.orghttps://www.mathematica.org/publications/evaluation-of-the-teacher-potential-project

Dougherty, C. (2016). *Keeping Track of Improvement in Educational Practices* (p. 24). ACT; University of Iowa.

Egalite, A. (2017, March 27). *The Effect of Louisiana's Voucher Program on School Integration: A Response to The Century Foundation*. Education Next. http://educationnext.org/effect-louisianas-voucher-program-school-integration-response-century-foundation/

Feng, L., & Sass, T. R. (2013). What makes special-education teachers special? Teacher training and achievement of students with disabilities. *Economics of Education Review*, *36*, 122–134. https://doi.org/10.1016/j.econedurev.2013.06.006

Fischer, C., Fishman, B., Dede, C., Eisenkraft, A., Frumin, K., Foster, B., Lawrenz, F., Levy, A. J., & McCoy, A. (2018). Investigating relationships between school context, teacher professional development, teaching practices, and student achievement in response to a nationwide science reform. *Teaching and Teacher Education*, *72*, 107–121. https://doi.org/10.1016/j.tate.2018.02.011

Hanushek, E. A., Piopiunik, M., & Wiederhold, S. (2019). Do Smarter Teachers Make Smarter Students? International Evidence on Teacher Cognitive Skills and Student Performance. *Education Next*, *19*(2). https://www.educationnext.org/do-smarter-teachers-make-smarter-students-international-evidence-cognitive-skills-performance/

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hendricks, M. D. (2014). Does it pay to pay teachers more? Evidence from Texas. *Journal of Public Economics*, *109*, 50–63. https://doi.org/10.1016/j.jpubeco.2013.11.001

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical Benchmarks for Interpreting Effect Sizes in Research. *CDEP Child Development Perspectives*, *2*(3), 172–177.

Holmes, W. M. (2014). *Using Propensity Scores in Quasi-Experimental Designs*. SAGE Publications, Ltd. https://doi.org/10.4135/9781452270098

Jackson, K., & Makarin, A. (2018). Can online off-the-shelf lessons improve student outcomes? Evidence from a field experiment. *American Economic Journal: Economic Policy*, *10*(3), 226–254.

Jimenez-Castellanos, O. (2010). Relationship Between Educational Resources and School Achievement: A Mixed Method Intra-District Analysis. *The Urban Review*, *42*(4), 351–371. https://doi.org/10.1007/s11256-010-0166-6

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, *27*(6), 615–631. https://doi.org/10.1016/j.econedurev.2007.05.005

Kane, T., Owens, A., Marinell, W., Thal, D., & Staiger, D. (2016). *Teaching Higher: Educators' Perspectives on Common Core Implementation* (p. 56). Center for Education Policy Research, Harvard University. http://cepr.harvard.edu/files/cepr/files/teaching-higher-report.pdf?m=1454988762

Lafortune, J., Rothstein, J., & Schanzenbach, D. W. (2018). School Finance Reform and the Distribution of Student Achievement. *American Economic Journal: Applied Economics*, *10*(2), 1–26. https://doi.org/10.1257/app.20160567

Lane, E., Linden, R., & Stange, K. (2018). *Socioeconomic Disparities in School Resources: New Evidence from Within-Districts*. University of Michigan. http://www-personal.umich.edu/~kstange/papers/LaneLindenStangeOct2018.pdf

Li, W., & Konstantopoulos, S. (2017). Does class-size reduction close the achievement gap? Evidence from TIMSS 2011. *School Effectiveness and School Improvement*, *28*(2), 292–313. https://doi.org/10.1080/09243453.2017.1280062

LIFT & TNTP. (2019). *Instructional Materials Implementation Guidebook*. TNTP. https://lifteducationtn.com/wp-content/uploads/2019/11/LIFT-Guidebook-2019-FINAL.pdf

NCES. (2006). *Appendix D - NCES Locale Codes* (p. 2). National Center for Education Statistics, Institute for Education Sciences, U.S. Department of Education.

Nichols-Barrer, I., & Haimson, J. (2013). *Impacts of Five Expeditionary Learning Middle Schools on Academic Achievement*. Mathematica Policy Research. https://www.mathematica.org/our-publications-and-findings/publications/impacts-of-five-expeditionary-learning-middle-schools-on-academic-achievement

Opfer, V. D., Kaufman, J., & Thompson, L. (2016). *Implementation of K–12 State Standards for Mathematics and English Language Arts and Literacy: Findings from the American Teacher Panel*. The Rand Corporation. https://www.rand.org/content/dam/rand/pubs/research_reports/RR1500/RR1529-1/RAND_RR1529-1.pdf

Parke, C. S., & Kanyongo, G. Y. (2012). Student attendance, mobility, and mathematics achievement in an urban school district. *The Journal of Educational Research*, *105*(3), 161–175.

Pham, L. D., Nguyen, T. D., & Springer, M. G. (2021). Teacher Merit Pay: A Meta-Analysis. *American Educational Research Journal*, *58*(3), 527–566. https://doi.org/10.3102/0002831220905580

Polikoff, M., & Koedel, C. (2017). *Big bang for just a few bucks: The impact of math textbooks in California | Brookings Institution* (Evidence Speaks #5). Brookings Institution. https://www.brookings.edu/research/big-bang-for-just-a-few-bucks-the-impact-of-math-textbooks-in-california/

Powell, M. G., Hull, D. M., & Beaujean, A. A. (2019). Propensity Score Matching for Education Data: Worked Examples. *The Journal of Experimental Education*. https://www.tandfonline.com/doi/full/10.1080/00220973.2018.1541850

Reardon, S. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In G. J. Duncan & R. J. Murnane (Eds.), *Whither opportunity?: Rising inequality, schools, and children's life chances*.

Reeves, P. M., Pun, W. H., & Chung, K. S. (2017). Influence of teacher collaboration on job satisfaction and student achievement. *Teaching and Teacher Education*, *67*, 227–236. https://doi.org/10.1016/j.tate.2017.06.016

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005a). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005b). Teachers, Schools, and Academic Achievement. *Econometrica*, *73*(2), 417–458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can You Recognize an Effective Teacher When You Recruit One? *Education Finance and Policy*, *6*(1), 43–74. https://doi.org/10.1162/EDFP_a_00022

Ronfeldt, M., & McQueen, K. (2017). Does New Teacher Induction Really Improve Retention? *Journal of Teacher Education*, *68*(4), 394–410. https://doi.org/10.1177/0022487117702583

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. https://doi.org/10.2307/2335942

Seah, K. K. (2018). Immigrant educators and students' academic achievement. *Labour Economics*, *51*, 152–169. https://doi.org/10.1016/j.labeco.2017.12.007

Shores, K., & Ejdemyr, S. (2017). *Pulling Back the Curtain: Intra-District School Spending Inequality and Its Correlates* (SSRN Scholarly Paper ID 3009775). Social Science Research Network. https://doi.org/10.2139/ssrn.3009775

Tibshirani, R. (1996). Regression shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society.*, *58*(1), 267–288.

TNTP. (2018). *The Opportunity Myth: What Students Can Show Us About How School Is Letting Them Down—And How to Fix It*. TNTP. https://tntp.org/publications/view/student-experiences/the-opportunity-myth

Tyler, J. H., Taylor, E. S., Kane, T. J., & Wooten, A. L. (2010). Using student performance data to identify effective classroom practices. *American Economic Review*, *100*(2), 256–260.

USDOE. (2016). *Non-Regulatory Guidance: Using Evidence to Strengthen Education Investments*. U.S Department of Education (ED), Office of Elementary and Secondary Education,. https://www2.ed.gov/policy/elsec/leg/essa/guidanceuseseinvestment.pdf

Wei, R. C., Darling-Hammond, L., Andree, A., Richardson, N., & Oprhanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the United States and abroad.* National Staff Development Council.

What Works Clearinghouse. (2020). *What Works Clearinghouse Standards Handbook, Version 4.1* (p. 116). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Appalachia. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf