# *Implementing Wit & Wisdom: An Evaluation*
## *Research Summary*
June 2021

*This study uses quasi-experimental propensity score matching to estimate the short-term impact of implementing Wit & Wisdom on students' reading outcomes in 1st through 5th grade. The first year of Wit & Wisdom implementation is associated with increases in students'* Text Reading and Comprehension (TRC) *scores in 1st through 3rd grade students and increases in students' ELA state test scores, on average.*

## *Research Overview*

Research shows that implementing a high-quality, knowledge-building curriculum can increase student learning and equity across classrooms,[1] while also providing support to teachers. Teachers are the single most important in-school determinant to students' success; however, it is the curricular materials teachers use that often shape *what* and *how* teachers teach.[2] While many factors contribute to strong academic achievement, adopting a rigorous curriculum can be an essential first step.

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),[3] 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 of inducing teachers to use standards-aligned curriculum effectively.

Quantitative research provides additional evidence of how long complete implementation can take: One three-year study of implementing a high-quality curriculum combined with professional learning[4] 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.

This study analyzes the impact of adopting the Wit & Wisdom (W&W) curriculum on student learning after the first year and provides preliminary estimates on the impacts of the curriculum.

During the 2018-19 academic year, 80% of elementary schools within a large suburban school district in North Carolina adopted W&W. The Johns Hopkins Institute for Education Policy ("Institute") analyzed this implementation through surveys, classroom observations, and analyzing administrative data. This study leverages administrative data to estimate the impact of the first year of implementation for 1st through 5th-grade students' literacy skills.

The Institute's findings should be interpreted as an analysis of the short-term impact of implementing the new curriculum. Given the research cited above, we might expect this impact to change over time.

## *Data*

In this study, we combine two sources of data. 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 report focuses on one measure of early literacy for 1st through 3rd grade students—the Text Reading and Comprehension (TRC) Benchmark Measure portion of the mCLASS: Reading 3D assessment—and the North Carolina End-of-Grade Reading test scores for 4th and 5th 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 that students need instructional support to fully comprehend and accurately read a book at an alphabetical reading level (i.e., "instructional"), or 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 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.

The North Carolina End-of-Grade Reading test is a multiple-choice test aligned to North Carolina state standards. In the elementary grades, the test is administered in 3rd through 5th 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 district that adopted W&W, and elementary schools within bordering districts. These data include information such as the school's size, locale, and the percentage of students who qualify for free and reduced lunch within each school.

### *School District Context*

The data come from a large suburban district in North Carolina that implemented the Wit & Wisdom curriculum during the 2018-2019 school year in 80% of its elementary schools. 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 over 100 elementary schools in the *matched* sample—the combined sample of schools using W&W and the nearby comparison schools—which provide education to a diverse student-body in settings ranging from remote rural to a midsized city.[5] Approximately 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 the intervention (i.e., W&W schools) and comparison schools (i.e., 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 |

Table 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 that 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.

## *Methodology*

This study uses quasi-experimental statistical techniques (i.e., ESSA Tier II evidence), and follows standards defined by WWC. We note that ESSA defines tiers of research evidence,[6] and does not require following procedures established by WWC. However, WWC standards can be adopted to establish evidence for ESSA.

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. Because this comparison is impossible, we use propensity score matching to create an apples-to-apples comparison. Specifically, we match students exposed to the intervention (i.e., students in schools using W&W) to students not exposed to the intervention (i.e., students in nearby schools not using W&W), and who are as identical to the intervention students as possible.

This matching method produces plausibly causal estimates. This means the results can be attributed to the intervention and not to other differences between the two groups under comparison. This Tier II study is more rigorous than a correlational Tier III study because we construct an observationally similar comparison group. This helps reduce bias and produces more accurate results. A Tier II study is the most rigorous research design possible for this curriculum implementation context since the curriculum was not randomly assigned to schools.

We match students on a variety of student characteristics, including grade, gender, race, pre-intervention test scores, disability status, ELL status, and age. In addition, we match on 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.[7] We also match on 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' 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.

An important aspect of the model for WWC guidelines is checking that the characteristics included in the model (i.e., covariates) were not impacted by the intervention status (i.e., endogenous). This model meets these WWC requirements because many measures were collected before the beginning of the intervention (e.g., 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). The WWC Standards Handbook clearly states that under these conditions, covariates are not potentially endogenous.[8]

WWC guidelines also 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. All grade levels in this report 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 |

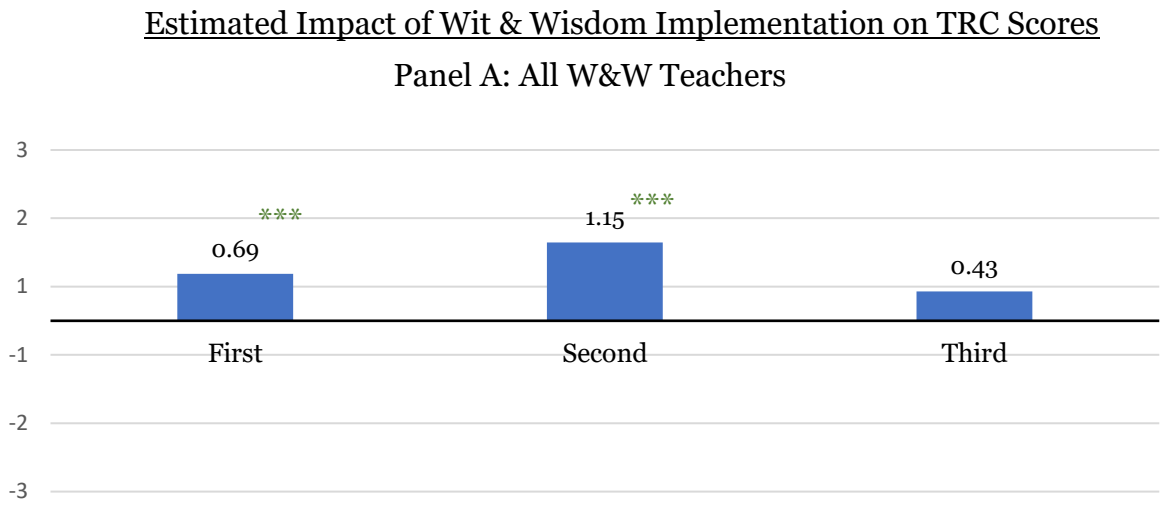
WWC guidelines also require adjustments to statistical significance when an intervention is adopted at a different level than reported estimates. In this study, the W&W intervention was adopted at the grade level (i.e., while most schools adopted the curriculum for all grades served in the school, some 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.[9] 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 statistical significance findings. These estimates follow an estimation procedure proposed by Abadie and Imbens[10] 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 Wit & Wisdom. 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.

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

*Figure 3*




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: ~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 1st through 3rd grade classrooms. These estimates are statistically significant in first and second grades, when we estimate statistical significance using Abadie and Imbens[11] estimators, but not 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 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.

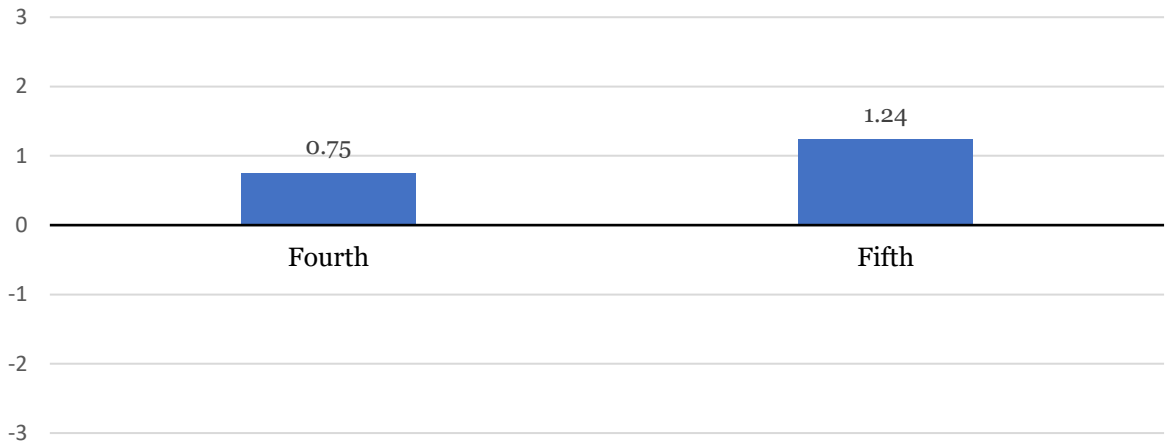
Issues of how to interpret the significance levels of the estimates aside, the results show that all students experienced better TRC scores than they otherwise would have, after their teachers switched to the W&W curriculum. For example, first-grade students whose teachers started using W&W read at almost 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 4th and 5th grade students' state test scores. These estimates show that both 4th and 5th 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 4th 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, 5th 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 estimates: ~p<.10, *p<.05, **p<.01, ***p<.001 and using WWC adjustments: ~p<.10, *p<.05, **p<.01, ***p<.001

## *Conclusions*

These results suggest that the implementation of 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 1st and 2nd grades are statistically significant when using standard estimation methods, but not under the adjustments required by WWC. Estimates in 3rd through 5th grades are positive, but statistically insignificant. Nonetheless, these estimates suggest that the implementation of W&W is associated with increased student literacy outcomes, even in the first year of implementation. These are promising findings given prior research on both the challenge of implementing new curriculum, and how long it often takes to fully realize the positive learning gains derived from implementing high-quality curricular materials.

[1] Rachana Bhatt and Cory Koedel, "Large-Scale Evaluations of Curricular Effectiveness: The Case of Elementary Mathematics in Indiana," *Educational Evaluation and Policy Analysis* 34, no. 4 (December 1, 2012): 391–412, https://doi.org/10.3102/0162373712440040; Rachana Bhatt, Cory Koedel, and Douglas Lehmann, "Is Curriculum Quality Uniform? Evidence from Florida," *Economics of Education Review* 34 (June 2013): 107–21, https://doi.org/10.1016/j.econedurev.2013.01.014.

[2] Steven G. Rivkin, Eric A. Hanushek, and John F. Kain, "Teachers, Schools, and Academic Achievement," *Econometrica* 73, no. 2 (2005): 417–58; Raj Chetty, John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *The American Economic Review* 104, no. 9 (2014): 2633–79; Thomas Kane et al., "Teaching Higher Educators' Perspectives on Common Core Implementation - Google Search," 2016, https://cepr.harvard.edu/files/cepr/files/teaching-higher-report.pdf.

[3] Sharon Roberts et al., "Instructional Materials Implementation Guidebook, 2019," 2019, https://lifteducationtn.com/wp-content/uploads/2019/11/LIFT-Guidebook-2019-FINAL.pdf.

[4] Ira Nichols-Barrer and Joshua Haimson, "Impacts of Five Expeditionary Learning Middle Schools on Academic Achievement" (Washington DC: Mathematica Policy Research, June 8, 2013), https://www.mathematica.org/our-publications-and-findings/publications/impacts-of-five-expeditionary-learning-middle-schools-on-academic-achievement.

[5] " NCES Handbooks- Appendix D NCES Locale Codes." National Center for Education Statistics, United States Department of Education, https://nces.ed.gov/programs/handbook/data/pdf/appendix_d.pdf Accessed 15 May 2021.

[6] ESSA sorts evidence into four categories or "tiers." The first tier provides "strong evidence" and is the most rigorous of the tiers but requires evidence supported by an experimental study. These studies can be difficult to field and therefore the top tier of evidence is difficult to obtain. The second ESSA tier of evidence provides "moderate evidence" supported by studies that approach the rigor of an experimental study, called "quasi-experimental" studies. The third tier of evidence provides "promising evidence" supported by well-designed correlational studies that control for some types of bias but are less rigorous.

[7] Department of State, "Teacher Program," *BridgeUSA*, accessed June 2, 2021, http://j1visa.state.gov/programs/teacher/.

[8] What Works Clearinghouse, "What Works Clearinghouse Standards Handbook, Version 4.1" (Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Appalachia, 2020), 33, https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf.

[9] What Works Clearinghouse, "What Works Clearinghouse Standards Handbook, Version 4.1" (Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Appalachia, 2020), F-2, https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf.

[10] Alberto Abadie and Guido W. Imbens, "Matching on the Estimated Propensity Score," *Econometrica* 84, no. 2 (2016): 781–807, https://doi.org/10.3982/ECTA11293.

[11] Abadie and Imbens.