NBER WORKING PAPER SERIES

PUTTING COMPUTERIZED INSTRUCTION TO THE TEST: A RANDOMIZED EVALUATION OF A "SCIENTIFICALLY-BASED" READING PROGRAM

Cecilia E. Rouse
Alan B. Krueger
with Lisa Markman

Working Paper 10315
http://www.nber.org/papers/w10315

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2004

We thank Jean Grossman, Rel Lavizzo-Mourey, and Rebecca Maynard for helpful suggestions, and the many dedicated principals, teachers and staff of the school district who implemented the Fast For Word Programs, provided data, and answered endless questions. We also thank Corinne Dretto, Sandy Ford, Suandra Hayward, Kathleen Hocker, and Steve Miller of the Scientific Learning Corporation for providing much insight into the program and answering many questions. We are also grateful to Maureen Bryne, Elizabeth Hester, Angela Oberhelman, Annabel Perez, and Kristen Russo who helped us to choose and administer the language test, and to Radha Iyengar and Alice Savage for expert research assistance. Finally we thank the Smith Richardson Foundation and the Education Research Sction at Princeton University for finacial support. All views and any errors are ours alone. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research.

Putting Computerized Instruction to the Test:
A Randomized Evaluation of a "Scientifically-based" Reading Program
Cecilia E. Rouse and Alan B. Krueger with Lisa Markamn
NBER Working Paper No. 10315
February 2004
JEL No. I2

ABSTRACT

Although schools across the country are investing heavily in computers in the classroom, there is surprisingly little evidence that they actually improve student achievement. In this paper we present results from a randomized study of a well-defined use of computers in schools: a popular instructional computer program, known as Fast ForWord, which is designed to improve language and reading skills. We assess the impact of the program using four different measures of language and reading ability. Our estimates suggest that while use of the computer program may improve some aspects of students' language skills, it does not appear that these gains translate into a broader measure of language acquisition or into actual reading skills.

Cecilia E. Rouse
Woodrow Wilson School
Robertson Hall
Princeton University
Princeton, NJ 08544-1013
and NBER
rouse@princeton.edu

Alan B. Krueger
Woodrow Wilson School
Princeton University
Princeton, NJ 08544-1013
and NBER
akrueger@princeton.edu

with the collaboration of Lisa Markman
Princeton University

I. Introduction

According to the 2000 *National Assessment of Educational Progress* (NAEP), 37 percent of 4th graders in the U.S. read below a basic level and an additional 31 percent read at a basic level, as determined by the National Assessment Governing Board (U.S. Department of Education, 2001). Currie and Thomas (2001) find that scores on a reading test taken at age 7 by participants in the *British National Child Development Study* are positively correlated with their earnings and likelihood of employment at age 33. Furthermore, adults who score higher on the literacy test in the *Adult Literacy Survey* have a greater probability of working and higher earnings if they do work (see, e.g., Sum, 1999). While the interpretation of the correlation between literacy and employment outcomes is unclear, it is very likely that improving literacy skills for troubled readers would generate important economic and social benefits.

Many children who have trouble reading actually have one or more learning disability that makes it difficult for them to benefit from traditional classroom teaching methods. Policymakers and educators have searched for alternative ways to help them. Because the parents and teachers of such students are often desperate to find effective approaches to improve reading skills, particularly in an era of high-stakes testing, a private market for education products has flourished. The proliferation of computers in schools has also helped fuel a market for educational software products. However, these educational products are often controversial and rarely evaluated using rigorous analytical methods. Rather, the customer – the school superintendent, technology officer or principal -- must often rely on research results produced and promulgated by the company itself, creating the potential for agency problems and conflicts of interest.

One popular, new product is a group of computer software programs known as the *Fast ForWord* (FFW) Family of Programs, distributed by Scientific Learning Corporation (SLC). Although FFW has only been available to public schools for about 5 years, it has already been used by over 120,000 students

and interest is growing. The number of students who used the program increased by 200 percent between 2000 and 2001 (Scientific Learning, 2002). These programs are based on the theory that many children with delayed development in language and reading have auditory processing difficulties (see Macaruso and Hook (2001) for a nice introduction to FFW). An example of an auditory processing disorder is when a child has difficulty distinguishing among consonant-vowel pairs such as /ba/ and /da/. The FFW programs attempt to retrain the brain to process information more effectively through a group of computer games that slow and magnify the acoustic changes within normal speech. As Macaruso and Hook (2001) explain, "...the [Fast ForWord] programs should help facilitate reading acquisition because they sharpen phonological processing skills ... which in turn benefit acquisition of phonic word attack strategies." (p. 6)

SLC claims that FFW generates impressive results. The best known studies underlying FFW are two articles published by highly respected neuroscientists in *Science* (Merzenich, et al, 1996; Tallal, et al., 1996). These papers report that language-learning impaired children participating in adaptive training exercises on the computer showed significant improvements in their "temporal processing" skills after 8-16 hours of training. These computer games evolved into the FFW programs. The FFW website (www.fastforword.com) highlights subsequent results based on national samples of children using the FFW language programs. Using data on about 1,200 students in grades K-6 in which some students were assessed using the Clinical Evaluation of Language Fundamentals (CELF-3) and others the Test of Language Development (TOLD) (both standardized tests of language skills), students showed gains of about 12 points. With a reported standard deviation of 15, these results suggest an effect size of 0.8F, which most educators would agree is an impressive impact. Further, data from about 300 students

(nationwide) in grades K-6 showed an increase of 3 points on the WJ-R (Woodcock Johnson Tests of Achievement, Revised), a test of actual reading skills, which suggests an effect size of approximately 0.2F. Paula Tallal, one of the researchers who founded SLC, claims, "After six to eight weeks, 90 percent of the kids who complete the program made 1.5 to two years of progress in reading skills...." (Begley and Check, 2000).

If these results represent the true effect of FFW on student language and reading development, then the program represents a remarkable addition to the tools available to schools to help students with reading difficulties. However, there are several reasons to suspect the results may be overly optimistic. First, many of the studies have very small sample sizes – for example, Tallal, et al. (1996) analyzed a sample of 22 children. Second, many of the results simply represent the difference in the test scores of students who participated in FFW from before the training and after the training. These students may have shown an increase in their language or reading skills even without the intervention due to their regular school instruction or maturation. One cannot determine whether the gains represent a causal effect of FFW because they are not compared to gains made by comparable students who did not participate in FFW over the same time period. That is, they lack a valid control group.

Researchers affiliated with SLC have conducted an evaluation of FFW that did employ a randomly selected control group.[1] Specifically, Miller, et. al. (1999) evaluate the effect of FFW on about 450 students from 9 elementary schools, drawn primarily from grades K-2. The students were evaluated on

[1] In addition, an unpublished independent randomized treatment-control evaluation of FFW has been conducted by Borman and Rachuba (2001). They find little impact of FFW on student achievement. We discuss their results in relation to ours in the conclusion.

three outcomes: the Test of Auditory Comprehension of Language, Revised Edition (TACL-R); the Phonological Awareness Test (PAT); and Single Word Reading (WJRWD) (Letter-Word Identification Subtest, Woodcock-Johnson Psycho-Education Battery-Revised). They report significant treatment effects for the FFW participants, both for the sample as a whole and for English-as-a-Second Language (ESL) students. A problem with this evaluation, however, is that the researchers appear to have excluded treatment students who did not complete the program, which may have introduced sample selection bias into their estimates. In addition, treatment students trained on FFW until they had "completed" the program as defined by Scientific Learning (see Section V, below), which raises the issue of whether the treatment students were tested after a longer time interval than the control students (which is not clear in the paper). In practice, many school districts implement the program such that students only train for a pre-specified period of time rather than to "completion" as defined by SLC.

In this paper we present results of a randomized evaluation of the FFW language programs. We initiated the evaluation with the help a superintendent of schools in a large urban school district in the northeast who wanted to know if FFW would improve the reading skills of children in his district before deciding whether to invest in the program on a large scale. After he invited us to study the program, we suggested using a within-school random assignment design to provide a control group that was otherwise similar to those selected for training on FFW. The time students spent using FFW was in addition to the amount of time they spent in regular reading instruction.

An important motivation for our evaluation is to provide new evidence on the potential impact of using computers in schools on student achievement. Available evidence on whether computers actually make a difference for students is quite small and the results are mixed (see Boozer, Krueger, and Wolkon,

1992, Angrist and Lavy, 2002, Wenglensky, 1998, Kirkpatrick and Cuban, 1998, and Goolsbee and Guryan, 2002). Because these studies are not based on random assignment, and except for Angrist and Lavy do not exploit a natural experiment, it is unclear whether omitted variables bias the estimated impact of computer use. Another limitation is that it is unclear exactly how the computers were used in the schools – that is, the nature of the treatment was not particularly clear or standardized. Many students may use computers with outdated or ineffective instructional software. *Fast ForWord* is the leading edge of scientifically-based computer technology in schools, and one of the more expensive programs available, so it provides a strong test. If students reap no benefit or only a small benefit from computer instruction with FFW, then it is unlikely that the average use of computers in schools generates any sizeable gains either.

Another motivation underlying our study is that claims for the success of FFW, and the development of the program itself, relied in large part on evidence derived from brain imaging (see, e.g., Temple, et al., 2000 and Nagarajan, et al. 1999). Scientific Learning Corporation's web page, for example, contains a link to a summary of a study of dyslexic children by Temple, et al. 2003, that reports that "activation of the children's brains fundamentally changed, becoming much more like that of good readers" after using FFW for 100 minutes a day for eight weeks. Advances in Functional Magnetic Resonance Imaging (fMRI) are affecting many areas of cognitive psychology, and even beginning to play a role in economics. Camerer, Lowenstein and Prelec (2003) provides a survey of ways in which fMRI technology has been used in economics research. An unresolved question, however, is whether responses detected in brain images translate into measurable changes in relevant skills and behaviors, such as reading ability. If not, then fMRI may be a tool that is of little more value to economics than is phrenology.

Although we were unable to implement brain scans for the particular subjects in our study, by using a rigorous random assignment procedure we can assess whether reading ability, as measured by a battery of standardized tests, was improved by an intervention that has been found to affect brain functioning in past studies.

Compared to control students, we estimate a small effect of being (randomly) assigned to train on FFW (that is statistically significant at the 10 percent level) on a composite score from a computer assessment known as *Reading Edge*. This assessment is designed to measure language and early reading skills and is sold by the SLC. However, we find no effect of being selected to train on FFW on language skills using the receptive portion of the CELF-3. And, while we estimate that those who received more comprehensive treatment (as reflected in the total number of complete days of training and whether or not the student had achieved the completion criterion recommended by the SLC) improved more quickly on the *Reading Edge* test, we estimate no such differential gain on the CELF-3. Finally, we find no statistically detectable effect of the program (among those selected to train on FFW and among those who complete the program) on reading skills as reflected in *Success-for-All* and the state standardized reading assessments. Overall our estimates suggest that while the FFW programs may improve a few aspects of students' language skills, it does not appear that these gains translate into a broader measure of language acquisition or into actual reading skills, at least as measured by commonly used standardized tests. However, our sample sizes were too small to detect *small* effects of FFW with statistical precision, so one could not rule out that the program had a small effect on student language and reading skills. Nevertheless, these impacts are much smaller than those promoted by the SLC.

The next section of the paper explains the FFW program and our evaluation more fully, section

three describes our data and section four our empirical strategy. We describe the results in section five and conclude in section six.

II. FFW and the Evaluation

1. The FFW Family of Programs

The FFW family of programs is comprised of three programs: FFW Language (elementary and middle school/high school versions), FFW Language-to-Reading, and FFW Reading. FFW language focuses on developing oral language skills that will create a "foundation for reading." The program focuses on four major areas that are deemed critical for language acquisition: phonological awareness, listening comprehension, language structures, and sustained focus and attention. (FFW Middle and High School Language contain much of the same content as FFW Language but have more mature graphics.) FFW Language-to-Reading focuses on making the connection between spoken and written language. The program attends to skills such as sound/letter recognition, decoding, vocabulary, syntax and grammar as well as listening comprehension and word recognition. FFW Reading focuses on building reading skills such as word recognition and fluency, decoding, spelling and vocabulary and passage comprehension.[2]

According to the FFW website, the target population for the program is "…anyone who wants to

---

[2] One subtlety of the FFW program is that once a student has achieved a particular level or if the student appears to have plateaued on his performance on particular subcomponents of the program, he should be advanced to the next program. Thus, for example, if the student is achieving well on FFW Language (elementary or middle/high school version), she should be moved to FFW Language-to-Reading. Similarly, a student may be moved from FFW Language-to-Reading to FFW Reading. (None of the students in this study transitioned to FFW Reading, although some did transition to FFW Language-to-Reading.)

improve language, reading and overall communication skills" including children who "struggle with basic language skills." Further, an article by Turner and Pearson, also cited on the SLC website, notes that "*Fast ForWord* Language candidates score below the normal range on standardized language test(s)." (www.fastforword.com) The training takes 6-8 weeks to complete during which the students work for 90-100 minutes per day, 5 days a week. All students begin at the basic level in each game, and progress to more advanced levels once they achieve a pre-specified level of proficiency. A student is deemed to have successfully completed the program once he or she has trained for at least 20 days (although 30 days is preferable) and completed at least 80 percent of a majority of the 5-7 games.

The programs cost about $30,000 for a one-year license for 30 computers, and the professional training package costs about $100 per site. In addition, the school must have computers with sufficient power to run the software, color printers, head phones, Y-connectors, a quiet place for the students to complete the program, and an adult who has received training to supervise the FFW students.[3]

2. The Setting of the Evaluation

We conducted the evaluation in an urban school district in the northeast. The district has a student enrollment of over 20,000 students; 40 percent of whom are African American and over 50 percent of

[3] The student need not use the program in a dedicated computer lab, but can use a computer in a classroom or in another place (such as a library). In two of the schools in this evaluation the students went to a dedicated computer lab to use the program; in the other two schools the students used it in the library (in one of the schools the use of the library was restricted to other students during that time; in the other they divided the library to accommodate the FFW class). Further, the students need not be supervised by a certified teacher; rather any adult who has received training on FFW would qualify. That said, the job of the adult supervisor can be quite demanding as the adult must know how to recognize students who are having difficulties and to devise strategies to help the students get back on track.

whom are Hispanic. Almost 70 percent qualify for the national school lunch program, and 56 percent speak a language other than English at home. The test score levels of the students in the district are well below average for the state and have been a great concern for the superintendent and others in the district. As a result, the representatives in the district are interested in programs that might help the students to succeed better in school. For example, all schools in the district have adopted a whole-school reform model. Most schools use *Success for All*; the others an alternative program.

We conducted the evaluation in four schools, designated A, B, C, and D. As shown in Appendix Table 1, these schools have high percentages of minority students and of students who are eligible for the national school lunch program. In addition, the percentage of students who speak a language other than English at home ranges from about 40 percent to almost 100 percent. Compared to the district as a whole these schools have a lower percentage of low-income students and a higher percentage of (potential) non-English speakers.

3. The Evaluation

In an attempt to parallel the target population for the FFW programs, we restricted the "eligible" population to students who scored in the bottom 20 percent (statewide) – or significantly below grade level – on the state's standardized reading test.[4] Using student scores on the state test from the 2001-2002 school year, we first identified eligible students for the evaluation. We then sent letters and consent forms

[4] Also as shown in Appendix Table 1, the average composite CELF-3 (NCE) score (on the receptive portion of the test) for the students in the evaluation was about 26, on a scale with a mean of 50, again indicating that the students were the target population for the program.

home to the parents of these students. At the same time, we asked the principals in the four schools to identify which students they believed would not be able to sit through 90-100 minutes of computerized instruction per day, which students had already transferred from the school between the time our lists were generated and the beginning of the evaluation, and which students were otherwise unavailable (such as the family was away on a long trip).[5] From the remaining list, we randomly selected students to participate in the FFW program (the "treatment" group); and the rest of the students comprised the control group. The randomization was done within each grade and school; in the analyses that follow we control for grade and school interactions (which we refer to as "randomization blocks or pools").

Although sales of FFW were originally targeted to professionals in private practice, the SLC has placed a growing emphasis on selling the product in public schools – in 2001 76 percent of its total revenues came from sales to public schools (Scientific Learning, 2002). Within a school setting, FFW is not generally targeted to an entire class, but rather to children with difficulties learning to read. As a result, it is often administered as a pull-out program (meaning that students are "pulled out" of their regular classroom instruction), or before or after school. And our evaluation was no exception. First, to generate an adequate sample size we had two groups of students using the program each day (delineated by grade level). Second, each school had to find a way to fit the training time into its unique schedule. Table 1 shows the subjects and activities that the FFW students missed during their training, and the percentage of the control sample that participated in each set of activities. FFW students trained during subjects such as

[5] The total number of eligible students in each "flight" (explained later in the text) and the reasons students did not participate in the evaluation are provided in Appendix Table 2; Appendix Table 3 shows the mean characteristics of the students who became part of the evaluation and those who did not.

homeroom, math, science, language arts and specials (art, music, gym). In addition, during the first flight, in one school students completed part of their training on FFW before or after school. In no case were students taken out of SFA. *The end result is that the counterfactual treatment received by the control students was mixed, but FFW was primarily an add-on to regular reading instruction.*

We bought 30 suitable computers for one school, and head phones, Y-connectors, and color printers for all of the schools (so that the instructors could properly track the students' progress using the SLC software which presents the data in color). We also bought year-long site licenses for 2 of the schools. We conducted site visits during both flights to ensure that the computer labs were properly set-up and that the teachers and instructors had been adequately trained. In general, representatives of the SLC were cooperative and provided much support. For example, the company provided training for the FFW instructors at the beginning of the evaluation, conducted periodic site visits, and provided telephone support throughout the evaluation. That said, the supervising teacher at one of the schools complained that the company was unresponsive to repeated requests for help with software failures.

III. Data

We assess the effect of FFW on four tests designed to reflect both language and reading skills. We have both pre- and post-tests for each outcome.

1. *Reading Edge* -- Accelerated Mode

The first outcome that we measure is from a computerized test called *Reading Edge*. This test was

recently purchased by the SLC.[6] According to the FFW website *Reading Edge* was "[d]eveloped by reading experts from Harvard, Stanford and Johns Hopkins to measure the language and early reading skills that are necessary for success." There are two versions of this test: the normal mode and the accelerated mode. The accelerated mode measures skills in the language areas (phonological awareness, decoding, and processing) that are associated with early reading. We chose to use the accelerated version because it is relatively shorter than the normal mode (35 minutes versus 60 minutes), and gathers reading-specific information. Although the assessment was designed for students in kindergarten through the second grade, the SLC website claims that it can "... also help measure the skills of older students who are having trouble learning to read." The assessment is administered on the computer, and treatments and controls took the test at the beginning and end of their flight.

The advantage of this test from our perspective is that it was easily administered to all of the students in the evaluation and provides a short-term assessment of whether FFW had a treatment effect (using an assessment that should be sensitive to the FFW program). The disadvantages are that: 1) students in the treatment group may perform better on the *Reading Edge* post-test simply because they have had more experience using computers; 2) although the *Reading Edge* manual refers to a validation sample of 350, no further information was given regarding reliability and validity, and it is not recognized as a valid language test; and 3) while the test is specifically designed to test the various aspects of the FFW program, it is unclear if the capabilities assessed in *Reading Edge* are associated with language acquisition and/or literacy skills.

---

[6] See *Reading Edge – Educator's Guide* (1999) for more information.

In theory the *Reading Edge* sample consists of all students in the evaluation (i.e., 512 students). However, 24 students were not administered a post-test and three others were missing components of the post-test, generating an analysis sample size of 485.[7] Within our sample, the mean of the *Reading Edge* (pre) test is 51 and the standard deviation 30.[8]

2. The Clinical Evaluation of Language Fundamentals - Third Edition, (CELF-3)

After consulting with numerous testing experts, we also chose to administer the receptive portion of the CELF-3 – Concepts and Directions, Word Classes, and Semantic Relationships – which we will refer to as the CELF-3-RP.[9] In addition, we administered the Listening to Paragraphs supplemental test of the CELF-3. The receptive portion of the CELF-3 measures how well students interpret word meaning (semantics), word and sentence structure (morphology and syntax) and recall and retrieve spoken language (auditory memory). Specifically this assessment requires individuals to interpret, recall and execute oral commands, perceive relationships between words, and interpret semantic relationships in sentences. The

---

[7] The results are identical if we include the three students missing parts of the *Reading Edge* post-test and include dummy variables indicating which components are missing. We exclude them so that the sample size is consistent across the components of the test.

[8] We attempted to obtain the mean and standard deviation for the *Reading Edge* for a more nationally representative sample from the SLC to no avail. Therefore, we use our in-sample standard deviation of 30 to estimate effect sizes which is likely an underestimate of the true standard deviation.

[9] Both the CELF-3 and the TOLD had been used by Scientific Learning in previous studies. When we consulted with language experts, the overall sentiment was that the tests were fairly equivalent. We chose to use the CELF-3 because the structure of the test allowed us to generate an overall receptive language score using only three subtests. Therefore, it required less time to administer and we were able to collect data on a larger sample. See *Clinical Evaluation of Language Fundamentals – Third Edition – Examiner's Manual* (1995) and *Clinical Evaluation of Language Fundamentals – Third Edition – Technical Manual* (1995) for more information.

three subtests that comprise the receptive portion of the CELF-3 do, however, have some visual cues. Therefore, we included the Listening to Paragraphs supplemental test in our battery because it assesses the comprehension, recall and interpretation of material presented orally without any visual prompts.

As these language tests are administered one-on-one and are therefore both disruptive to the school and expensive, we chose to only administer them to a random sample of students in the fourth grade. We calculated that with 70 students we could detect a treatment effect of 0.4 of a standard deviation, which is smaller than Scientific Learning's reported effect size of about 0.8 of a standard deviation on language tests, including the CELF-3. We used three certified evaluators (who were independent of the district) to administer the tests and these evaluators did not know which students had been chosen for the treatment group and which were part of the control group. These tests were administered to a random sample of both the treatment and the control groups at the beginning and end of the second flight, and the same evaluator evaluated the same child both times. We present estimates that have been transformed to the normal curve equivalent with a mean of 50 and a standard deviation of 21.06. We managed to administer the post-test to all of the students (randomly) selected for the CELF-3-RP, generating a sample size of 89.

3. *Success For All* Assessments

Although language skills are critical to reading and other aspects of student performance, the district's ultimate concern is whether the FFW intervention improves students' reading skills. We assess the impact of FFW on students' reading skills using the 5 assessments administered throughout the year as part of the *Success For All* (SFA) whole-school reform model. SFA is a highly structured program that provides 90 minutes of uninterrupted daily reading to students who are grouped according to their reading

level regardless of their classes and/or grades. The program provides schools with curricula, assessment tools, and professional development, as well as tutoring and family support approaches. The SFA reading curriculum focuses on providing students with a balance of both phonics and meaning.

The assessments, which are closely aligned to the SFA curriculum, are given every eight weeks by the reading teachers.[10] The SFA facilitator from each school collects and interprets this information in order to suggest changes in grouping, tutoring, and/or classroom reading approaches. The assessments include both a paper-and-pencil assessment as well as a more subjective assessment by the student's teacher. The subjective SFA score is a combination of the test score and the student's class work during the 6 week interval. Further, students may be administered one of four versions of the assessments. "Reading Roots" (or "Roots") is administered in English to students with kindergarten and first-grade level reading skills. "Reading Wings" (or "Wings") is administered in English to students who read at at-least the second grade level. In addition, there are two versions of the test that are administered in Spanish ("Lee Conmigo" which corresponds "Roots" and "Alas" which corresponds to "Wings") for those with limited English proficiency. Although the different versions have different score scales, we converted them all to the "standard method" (that used by Wings) that ranges from 1.1-9.9. In principle, a score of 1.1 means that the student is reading at the first grade, first month level (approximately October); a score of 3.5 means that the student is reading at the third grade, fifth month level. In the SFA data we obtained from the district, the average score on the initial assessment was 3.7 with a standard deviation of 1.5.

One disadvantage of the SFA data is that part of the student's score reflects the SFA evaluator's

[10] The initial assessment is administered in September, the first in November, the second in January, the third in March, and the final in June.

subjective assessment, and the evaluator (who is not the student's regular teacher) may have been aware of which students were assigned to the FFW treatment group and which were part of the control group. An advantage of the SFA data is that the assessments coincided nicely with the beginnings and ends of our FFW flights. Also, the SFA assessments more closely reflect outcomes that the superintendent, principals and teachers care about, such as reading and writing achievement, and educational behaviors and habits (e.g., note taking, direction following, attention and focus).

Our sample includes students from 17 regular elementary schools in the district. (We excluded 9 schools because the SFA data (which are collected by the individual schools) were deemed too unreliable by the district.) Of the original 512 students in the evaluation, 124 of them attended a school that does not use SFA, 10 others were simply not in the SFA data files, and 4 others were missing the SFA outcome variable. This leaves a sample of 374 (197 are treatments and 177 are controls).

4. State Standardized Reading Tests

Our final outcome is the student's score on the state's criterion-referenced standardized test. The exam is designed to be aligned with the curriculum standards of the state as well as to parallel critical aspects of the *National Assessment of Educational Progress* (NAEP). The state administers tests in reading, math, and writing to 4^{th}, 6^{th}, and 8^{th} graders annually. The district in which we conducted the evaluation also conducts "off-year" tests to 3^{rd}, 5^{th}, and 7^{th} graders. The district attempts to score the off-year tests to match the state scoring procedures for the "raw scores." (The district is unable to construct "scale scores" for the off-years because it is one of only a few districts that administer the "off-year" tests.) We translate the total raw scores into "district" percentile scores in order to have a measure that is readily

interpreted. We do so by constructing (district) percentiles using the scores of students not enrolled in schools that participated in the FFW evaluation for each subject and year. We then determine the district percentile of the scores for students in the schools in the evaluation.[11] Because we use percentile scores, which are uniform in the district, the standard deviation is 28.9.

Our strategy with the state test sample is similar to that using the SFA data. Of the students in the original evaluation, 58 were missing follow-up state test data (mostly likely because they had transferred out of the district) leaving a sample of 454 (237 are treatments and 217 are controls).

5. Correlation Among Outcomes

The four assessments are relatively highly correlated, especially among the more established tests. For example, the state's reading, writing, and math tests show correlations of about 0.73, which is common among nationally normed tests.[12] More importantly, the CELF-3-RP (which is a language test) and *Reading Edge* (which is a combination language and reading skills test) show correlations of about 0.2-0.4 with the state's reading assessment; the SFA assessments (which are reading tests) show correlations of over 0.8 with state reading test. All three external assessments also show similar correlations with the state's math assessment. These correlations suggest that although the CELF-3-RP and *Reading*

---

[11] Although we present results here using the district percentile scores, the results are nearly identical if we use the total raw scores.

[12] For example, Krueger (1997) reports that in the Tennessee class size experiment (known as project STAR) the correlation between the math and reading portions of the Stanford Achievement Test was 0.69 for first-graders and 0.73 for second graders. Similarly, in the New York City School Choice Scholarships Program (a school voucher experiment) the correlation between the Iowa Skills Test for the control students in grades 3-6 was 0.66 in the third year (authors' calculations).

Edge outcomes are designed to assess the building blocks of reading, they are also somewhat related to actual reading skills. Importantly, the SFA assessments are strongly correlated with the state's reading assessment and therefore a second barometer of the effect of the FFW programs on reading.

IV. Empirical Strategy

We evaluate the program using two statistical techniques. First, we estimate models that generate estimates of the "intent-to-treat" effect using FFW. In these models, the test scores of students randomly assigned to participate in FFW are compared to the scores of students randomly assigned to the control group, whether or not the students remained in their original assignments. (That is, if a student was assigned to the treatment group but did not actually participate in FFW or did not train for many hours, we still count the student as having participated. Similarly, if a student was assigned to the control group but ultimately participated in FFW we still consider the student a control.) An ordinary least squares (OLS) regression of the following model generates an unbiased estimate of the intent-to-treat effect:

$$Y_i = \alpha + X_i\beta + \gamma T_i + \delta P_i + \varepsilon_i \qquad (1)$$

where Y_i represents student *i*'s score on one of the follow-up tests, T_i indicates whether the student was randomly selected to participate in FFW, X_i represents a vector of student characteristics (including the student's baseline test scores), P_i represents the pool from which the student was randomly selected (an interaction between the student's grade and school), and ε_i is a random error term; ", $, (, and * represent coefficients to be estimated. We present estimates both with and without the vector X_i.

The coefficient (represents the "intent to treat" effect and estimates the effect of assigning a student to the treatment group on the outcome in question. While the intent-to-treat effect represents the gains that an educator can realistically expect to observe from implementing the program (since one cannot fully control whether students actually participate in the program or train for the required number of hours), it does not necessarily represent the effect of the program for those who actually complete it.

Therefore, we also estimate instrumental variables (IV) models in which we use a dummy variable indicating whether the student was randomly selected to participate in FFW as an instrument for actual participation. The random assignment is correlated with actual participation in FFW but uncorrelated with the error term in the outcome equation (since it was determined randomly). Under plausible assumptions, this model yields a consistent estimate of the effect of "treatment on the treated." In this case, the second-stage (outcome) equation is represented by models such as,

$$Y_i = \alpha'' + X_i\beta'' + \lambda FFW_i + \delta'' P_i + \varepsilon_i'' \quad . \qquad (2)$$

FFW_i is one of three measures of whether the student actually participated in FFW. The first is the total number of complete days the student trained on FFW. The second and third measures reflect whether the student completed the FFW training as determined by a draft Scientific Learning protocol: whether the student completed at least 20 days of training and completed at least 80 percent of a majority of the games; and whether the student completed at least 30 days of training and completed at least 80 percent of a majority of the games.[13] δ indicates the effect of participation/completion in FFW on student outcomes,

[13] In both cases we only identify a student has having completed a majority of exercises if he or she has completed at least 80 percent (or 90 percent) of at least one sound exercise and at least one word exercise.

and the other variables and coefficients are as before. Through the use of IV one can generate a consistent estimate of the effect of FFW on student outcomes.

V. Results

1. Descriptive Statistics

To begin, we examine whether the treatment and control groups appear similar prior to random assignment. While balance among observable measures cannot ensure comparability of treatment and control groups, it can at least lend some suggestion of whether the randomization was successful. Table 2 shows the means and standard deviations of pre-treatment measures. The table shows that the mean differences between the treatments and controls are not statistically significant at traditional significance levels for all measures but the state writing score, which likely occurred by chance.

One problem with many evaluations is that the researchers have follow-up data on relatively small percentages of the original students. However, since we measure the effect of the program using short-term assessments and have data from the entire district for the SFA data and for the 2002-2003 state tests (which allows us to observe the test scores of students who transfer schools), this analysis does not suffer from large attrition of students. For example, we have a post-FFW CELF-3-RP test score for all of the sampled students; a post-Reading Edge test score for 95 percent of both treatments and controls; a post-SFA assessment for 96 percent of treatments and 97 percent of controls in the three SFA schools in the evaluation; and a state assessment in the 2002-2003 school year for 87 percent of treatments and 90 percent of controls.

Contamination can occur in randomized evaluations if control students nonetheless receive

treatment (through another channel) or if treatment students do not complete the program. In this evaluation, four of the control students actually trained on FFW. In addition, 8 students in the treatment group either transferred out of the school or never showed show up for training. Another important issue is whether the students who were selected for the program actually completed treatment.

Completion of FFW is a function of the amount of training time and whether the student actually makes progress on the program as reflected in the percentage of exercises that she has mastered.[14] Therefore, according to SLC's guidelines, a student has "completed" FFW if she has:

a) achieved at least 90 percent completion on all exercises regardless of the number of days of training; or,

b) trained for a minimum of 20 days (although an average of 30 days is preferred); and has at least 80 percent completion on a majority of exercises; and has shown steady progression in both sound and word exercises (not only in one or the other).[15]

We implement these guidelines for three measures of completion of the FFW "treatment." The first is the total number of "complete" training days. For a student to "complete" a day, he or she must train for 60 minutes for the first three days, 80 minutes for the fourth and fifth days, and 100 minutes thereafter for FFW Language. The student must train for 90 minutes a day for FFW Middle School and Language-to-Reading. Note that for this measure we do not count those days during which a student trains for fewer

[14] Recall that a student does not progress to the next level of a program until she has attained a pre-specified level of proficiency at the current level.

[15] The recommended training guidelines from SLC also suggest that if a student has not reached at least 80 percent in a majority of exercises but a majority of the exercises is at a plateau for at least 7 training days (plus the other two criteria) then the student should be considered finished. We only have the percent completion data for the last day of training and therefore cannot easily incorporate the students' trends over time.

than the required number of minutes for that particular day.[16] The second and third measures also incorporate a concept that SLC refers to as "percent complete," or the percentage of exercises that a student completes. For this measure we assess whether the student has completed at least 80 percent of a majority of the exercises (where at least one of these exercises is a word exercise and at least one is a sound exercise). Thus, for the second treatment measure we require that a student has completed at least 20 days of training and that she has completed at least 80 percent of a majority of the exercises. For the third measure we require that a student has completed at least 30 days of training as well as at least 80 percent of a majority of the exercises.[17]

As shown in Table 3, among those who did train, 76 percent in the first flight and 67 percent in the second completed at least 20 days of training (the minimum required for successful completion of FFW). However, only 51 percent of those who trained in the first flight and 38 percent in the second both completed the requisite amount of training and completed a sufficient fraction of the exercises. Although we believe that from the schools' perspective the implementation went more smoothly in the second flight than in the first, a smaller proportion of the students actually appear to have completed the program. We suspect that this is mostly due to the fact that the students in the second flight had fewer potential training days – 30, on average, compared to 37 for the first flight.

---

[16] In personal communication with representatives of SLC, they also only intend that "complete" days be counted when determining the number of days that a student has trained.

[17] We did not attempt to analyze the guideline that states that once a student has completed at least 90 percent on all of the exercises she has completed her training since only 9 students completed the program by that criteria.

2. Effects of Intent-to-Treat and Treatment-on-the-Treated

A. Effects of Intent-to-Treat

Table 4 presents the basic difference-in-differences estimates for the four test score outcomes. Note that in this table we do not control for baseline characteristics, nor do we control for the randomization pool from which the student was randomly drawn. Among the students selected to train on FFW, we estimate educationally large and statistically significant test scores gains for the *Reading Edge*, CELF-3-RP, SFA, and state standardized reading outcomes. The students posted a 21 point gain on the *Reading Edge* (which represents 0.7F), a 6.3 point gain on the CELF-3-RP (which represents 0.3F), a 0.27 point gain on the SFA (which represents 0.18F), and a 5.7 percentile point gain on the state reading test (which represents approximately 0.2F). All four effect sizes would be considered relatively large among educational interventions.

However, the students in the control groups posted nearly identical gains during the same period. Specifically, the controls gained 17.7 points on the *Reading Edge*, nearly 6 points on the CELF-3-RP, 0.25 points in SFA, and 4.4 percentile points on the state standardized reading test. As a result, the difference-in-difference (or treatment-control) estimate of the effect of FFW on student language outcomes is small and statistically insignificant on the *Reading Edge,* and virtually zero in the CELF-3-RP, SFA, and state standardized reading tests.

The estimates in Table 4 do not control for the student's randomization pool nor for any other covariates; these results are presented in Tables 5a and 5b.[18] Table 5a presents results for *Reading Edge*

[18] The sample size for the *Reading Edge* results in Table 5a include those students who were missing all or some of the components of the *Reading Edge* pre-test. We include dummy variables

and the CELF-3-RP; Table 5b presents results for the SFA and state standardized reading test. Consider first Table 5a. In columns (1) and (4) we only control for the student's randomization pool; in columns (2) and (5) we add the corresponding pre-test score; and in columns (3) and (6) we add student sex and race/ethnicity. The basic estimate for the *Reading Edge* test in column (1) is smaller than that in Table 4 and not statistically significant. The basic estimate for the CELF-3-RP in column (4) is similar to that in Table 4 – there is no detectable effect of being selected for FFW on the composite CELF-3-RP score. For the *Reading Edge* test, controlling for the pre-test increases the estimated intent-to-treat coefficient such that in columns (2) and (3) the effect is just barely statistically significant at the 10 percent level. Given a standard deviation of about 30 on this test in our sample – which is undoubtedly an underestimate of the standard deviation in the population – the estimates suggest an effect size of about 0.1F. The results using the CELF-3-RP remain small and statistically insignificant regardless of whether or not one controls for covariates.[19]

Table 5b presents estimates of the intent-to-treat effect using the SFA and state standardized reading assessments. Recall that while the *Reading Edge* and CELF-3-RP should reflect the student's language and early reading skills, the SFA and state standardized reading assessments should reflect whether such building blocks translate into actual reading gains. The estimates in columns (1)-(3) parallel those in Table 5a. In Table 5b, controlling for the student's randomization pool increases the estimated

---

indicating which indicate if a component is missing.

[19] While it does not matter whether one controls for covariates or not for the composite score on the CELF-3-RP, we estimate a large intent-to-treat effect with a p-value of 0.105 on the Listening to Paragraphs component of the CELF-3-RP before we control for the pre-test; with the pre-test this effect decreases in magnitude and becomes statistically insignificant. These results are available from the authors on request.

effect on the SFA assessment from that in Table 4 from 0.02 to 0.07. However, the estimate in Table 5b is still statistically insignificant and represents an effect size of only 0.05F. Adding covariates in columns (2) and (3) lowers the intent-to-treat estimate and improves the precision. The results in columns (4) - (6) for the state standardized reading test are similar: the coefficient estimates suggest an effect size of about 0.04-0.06F, although the point estimates are not close to statistically significant, with t-ratios less than 1.0. Based on these results, we conclude that, overall, there was no detectable effect of the FFW program on the reading skills of students.

In addition, we have estimated the intent-to-treat effect on all four assessments for various subgroups of students.[20] We find that, overall, being selected for FFW had a statistically significant (at the 10 percent level) effect on *Reading Edge* for girls but not boys, and that the gains were larger in the first flight than in the second. More generally, however, and especially for the other three assessments, we estimate no statistically detectable differences among subgroups of students.

In sum, we estimate a small effect of the being (randomly) selected to train on FFW (that is statistically significant at the 10 percent level) on the *Reading Edge*, however we estimate no statistically detectable effect of being selected to train on FFW on language skills using the CELF-3-RP or on reading skills using SFA and state standardized reading assessments.

B. Effects of Treatment-on-the-Treated

While intent-to-treat estimates are critical to good policy, they do not indicate whether FFW is

[20] These results are available from the authors on request.

effective for those who show up and actually train on FFW as well as for those students who complete the program as advised by the SLC – the effect of treatment-on-the-treated. We generate these estimates using IV models based on equation (2), as presented in Table 6. As noted above, we define treatment as the total number of complete training days, whether the student has completed at least 80 percent of the exercises and completed either at least 20 or 30 days of training.[21]

The results for the *Reading Edge* assessment are consistent across the three measures of having received treatment: students who receive more treatment improve more on the composite measure from *Reading Edge* than the control students, at a statistical significance level of 10 percent. These gains are concentrated in the non-word recognition, and to a lesser extent phoneme blending, sub-components of the test. In contrast, we continue to estimate no statistically detectable effect of the program on the CELF-3-RP, SFA, and state standardized reading assessments, although we also would not reject that those who actually trained for at least 30 days on the FFW program showed an improvement of 0.1F on the state standardized reading test.[22]

VI. Conclusion

This paper presents results from a randomized study of a well-defined use of computers in schools:

---

[21] Although we do not present them, the first stage equations suggest that whether a student is randomly selected to train on FFW is a strong predictor of the student's total completed days of training and on the two completion criteria. These results are available from the authors on request.

[22] The effect sizes for those who trained at least 30 days and completed at least 80 percent of the exercises suggest larger effects of actually completing the program (for the example the effect sizes are 0.17F for the CELF-3-RP, 0.08F for the SFA, and 0.15F for the state reading assessment.) However, if there is some gain to the program for those students who train on FFW but do not actually complete the program these estimates will be upward biased (see, e.g., Rouse 1997).

a popular instructional computer program designed to improve language and reading skills. Our estimates suggest that while the FFW programs may improve some aspects of students' language skills, it does not appear that these gains translate into a broader measure of language acquisition or into actual reading skills. However, one could be concerned that our sample sizes were too small to detect small effects of FFW with sufficient precision. Therefore, it is worth assessing the program by considering confidence intervals for the intent-to-treat effect sizes. The 95 percent confidence interval for the CELF-3-RP, a well-known language assessment, indicates that selection to train on the FFW programs generated an average gain of between -0.24F and 0.31F. Because of the larger sample size, we obtained a tighter interval for the effect of FFW on reading skills, as assessed by the state reading assessment; the 95 percent confidence interval for the effect of selection to train on FFW for this outcome covers effect sizes from -0.08F to 0.16F. We estimate larger effect sizes if we adjust for program non-completion, although the confidence intervals are somewhat wider and non-completion is a fact of life when the program is implemented in actual classroom settings.

It is useful to compare our results to those reported by Borman and Rachuba (2001), the only other randomized evaluation of FFW that was conducted independently of the SLC. Their study was implemented in 8 Baltimore City Public Schools and consisted of 415 children in grades 2 and 7; their target population was those students who scored below the 50th percentile on the Total Reading components of the *Comprehensive Test of Basic Skills*, 5th Edition (CTBS/5). Selected students trained on FFW for up to 8 weeks. Their 95 percent confidence interval for the intent-to-treat effect of being selected to train on FFW on the reading test runs from -0.08F and 0.13F – a range that is remarkably close to the one we estimate for the state reading assessment in our sample.

In both studies, large impacts of the computerized instruction can clearly be ruled out. But the potential benefits must be weighed against the costs. The direct costs of the program are not particularly high. A year-long site license costs about $30,000; adding in the cost of computers, printers, and other hardware leads to a total cost of software and hardware of about $37,000 per school per year for 20 stations. In addition one must have a trained adult with the students – assume a salary of $55,000 (the average teacher salary in the state). If the adult can supervise 40 students per day for 3 rounds of FFW training during the school year, this generates a cost (excluding the cost of the space) of about $770 per student. As a result, implementing FFW may be cost-effective for school districts with many students who would be appropriate for the FFW training and who are eager to help such students improve their language and early reading skills. Nevertheless, the direct cost of the program may be swamped by the indirect costs incurred as schools juggle schedules to accommodate the 90-100 minutes per day required by the program and provide a qualified adult to supervise the program – all for relatively small academic gains by students.

In any event, results from our experimental evaluation, along with those in Borman and Rachuba (2001), suggest that the achievement gains schools can expect students to experience from the FFW program are likely much smaller than those claimed by the vendor of the program.[23] In addition, the results suggest that the disappointing results on the impact of computers on student achievement that have been

[23] While we suspect that the claims made by SLC are inflated because researchers neglected to compare the gains of treated students to those of a control group, and focused only on those who successfully progress through FFW, we also suspect that the program may have been less effective for participants in our study because students had a surprisingly difficult time completing the program as recommended by the SLC. Further, we suspect that this difficulty may be somewhat inherent to implementing FFW in an urban school setting, as Borman and Rachuba (2001) encountered similar problems.

reported in the previous literature may not solely be due to the fact that the use of the computers was not well defined or state of the art, or to the lack of randomly selected treatment and control groups in those studies. Rather, it may be because computers are not an effective substitute for traditional classroom instruction, or because educators have not learned how to effectively use computer technology to enhance instruction, or because there are other aspects to the school setting that make it difficult to incorporate computerized instruction into the curriculum.

References

Angrist, Joshua and Lavy, Victor. "New evidence on classroom computers and pupil learning." *The Economic Journal*, no.112, October 2002, pp.735-765.

Begley, Sharon with Erika Check. "Rewiring Your Gray Matter." *Newsweek*, January 1, 2000, p. 63.

Boozer, Michael, Alan B. Krueger and Shari Wolkon. "Race and School Quality Since *Brown vs. Board of Education*," *Brookings Papers on Economic Activity: Microeconomics*, Martin N. Bailyand Clifford Winston (eds.), 1992, pp. 269-326.

Borman, Geoffrey D. and Laura T. Rachuba. "Evaluation of the Scientific Learning Corporation's Fast ForWord Computer-Based Training Program in the Baltimore City Public Schools." *A Report Prepared for the Abell Foundation*, August, 2001.

Camerer, Colin, George Lowenstein and Drazen Prelec. "Neuroeconomics: How Neuroscience can Inform Economics," Forthcoming, *Journal of Economic Perspectives* 2003.

Clinical Evaluation of Language Fundamentals - Third Edition -Examiner's Manual. The Psychological Corporation, San Antonio, TX, 1995.

Clinical Evaluation of Language Fundamentals - Third Edition - Technical Manual. The Psychological Corporation, San Antonio, TX. 1995.

Currie, Janet and Duncan Thomas. "Early Test Scores, Socioeconomic Status, School Quality and Future Outcomes," *Research in Labor Economics*, vol. 20, 2001.

Goolsbee, Austan and Jonathan Guryan. "The Impact of Internet Subsidies in Public Schools." NBER Working Paper No. 9090, August 2002 .

Hotz, Robert Lee. "Discovery: New Techniques Let Researchers Observe Neural Activity as Children Read. Understanding How the Mind Works Could Reshape Classroom Instruction." *Los Angeles Times*, October 18, 1998.

Kirkpatrick, Heather and Larry Cuban. "Computers Make Kids Smarter – Right?" *Technos Quarterly*, 70, no. 2 (Summer 1998).

Krueger, Alan B. "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, NBER Working Paper No. 6051, June 1997.

Macaruso, Paul and Paula E. Hook. "Auditory Processing: Evaluation of Fast ForWord For Children with Dyslexia." *Perspectives*, vol. 27, no. 3 (Summer 2001), pp. 5-8.

Merzenich, Michael M., William M. Jenkins, Paul Johnston, Christoph Schreiner, Steven L. Miller, and Paula Tallal. "Temporal Processing Deficits of Language-Learning Impaired Children Ameliorated by Training." *Science*, vol. 271 no 5 (January 1996), pp. 77-81.

Nagarajan, Srikantan, Henry Mahncke, Talya Salz, Paula Tallal, timothy Roberts, and Michael Merzenich. "Cortical auditory signal processing in poor readers," *Proceedings of National Academy of Sciences*, vol. 96, May 1999, pp. 6483-88.

Reading Edge - Educator's Guide 1999. Scientific LanguageCorporation. Berkeley, CA. 1999.

Rouse, Cecilia Elena. "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program," NBER Working Paper Number 5964 (March, 1997).

Scientific Learning Corporation. "2001 Annual Report." Oakland, CA. March, 2002.

Sum, Andrew. "Literacy in the Labor Force: Results from the National Adult Literacy Survey." U.S. Department of Education, National Center for Education Statistics, Washington, D.C., NCES 1999-470, September 1999.

Tallal, Paula, Steve L. Miller, Gail Bedi, Gary Byma, Xiaoqin Wang, Srikantan S. Nagarajan, Christoph Schreiner, William M. Jenkins, and Michael M. Merzenich. "Language Comprehension in Language-Learning Impaired Children Improved with Acoustically Modified Speech." *Science*, vol. 271 no 5 (January 1996), pp. 81-84.

Temple, Elise, R.A. Poldrack, A Protopapas, S. Nagarajan, T. Salz, P. Tallal, and M. Merenich, "Dirsruption of the neural response to rapid aoustic stimuli in dyslexia: Evidence from functional MRI," *Proceedings of the National Academy of Sciences Early Edition*. Vol. 97, no. 25, December 5, 2000, pp. 13907-12.

Temple, Elise, Paula Tallal, John Gabrieli, Gayle K. Deutsch, Russell Poldrack, Steven L. Miller, Michael M. Merzenich "Neural deficits in children with dyslexia ameliorated by behavioral remediation: Evidence from functional MRI," published by the *Proceedings of the National Academy of Sciences Early Edition*. 2003.

Turner, Shannon and Donise W. Pearson. "Fast ForWord Language Intervention Program: Four Case Studies," *Tejas: Texas Journal of Audiology and Speech Pathology*, vol. 13 (Spring/Summer 1999).

U.S. Department of Education. "The Nation's Report Card: Fourth Grade Reading Highlights, 2000," Washington, D.C. 2000.

Wenglinsky, Harold. "Does it compute? The relationship between educational technology and student achievement in mathematics." Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. ED425191) 1998.

Table 1:
The Subjects/Activities Missed During FFW Training, by School and Grade

School	Percentage of Control Sample *	Subject/Activities Missed During FFW Training
		Grade 3
A	5.4%	Specials, after school
B	8.7%	Some math, writing, language arts, spelling, science
C	8.7%	Science, social studies
D	5.4%	Math, some lunch, writing, language arts, science, social studies
		Grade 4
A	9.2%	Spelling, character education, language arts, and reading
B	6.3%	Language arts, extra reading, spelling
C	7.1%	Math, language arts, writing, social studies, science
D	7.1%	Writing, language arts, social studies, science
		Grade 5
A	7.9%	Before school, some homeroom, some language arts
B	5.0%	Specials, some math
C	7.9%	Homeroom, some language arts
D	2.5%	Writing, some social studies, language arts, math
		Grade 6
A	5.0%	Language, math, science, and specials
B	4.2%	Math, language arts, social studies, science, literature, grammar, some lunch
C	5.4%	Homeroom, some language arts
D	4.2%	Specials, lunch, math, recess

Notes: Specials include arts, music, gym, etc. In most cases during the time of the FFW training there was not a regularly scheduled subject. Thus, if (for example) math is among the subjects listed it does not imply that the FFW student only missed math instruction, but rather the range of subjects.

* These percentages reflect the percentage of the control sample that participated in each set of activities while the treatment students trained on FFW.

Table 2:
Pre-Selection Means and Standard Deviations
and Differences Between Treatment and Control Students

	FFW	Control	p-value of difference*
Female	0.489 [0.501]	0.546 [0.499]	0.176
African American	0.268 [0.444]	0.267 [0.443]	0.876
Hispanic	0.640 [0.481]	0.671 [0.471]	0.585
Identified as a Special Education Student	0.147 [0.355]	0.163 [0.370]	0.552
Pre-Composite State Reading Score	38.21 [19.52]	37.94 [18.69]	0.739
Pre-Composite State Writing Score	41.23 [23.31]	45.80 [24.41]	0.037
Pre-Total State Math Score	44.45 [26.40]	47.86 [26.60]	0.155
Pre-Composite CELF-3-RP Score	25.74 [18.75]	25.43 [18.03]	0.951
Pre-Composite *Reading Edge* Score	50.77 [30.65]	36.16 [23.96]	0.189
Pre-SFA Assessment	3.83 [1.38]	3.78 [1.36]	0.663

Notes: Standard deviations in brackets. The state test scores are percentiles from the 2001-2002 school year. CELF-3-RP scores are normal curve equivalents. There are a maximum of 272 FFW students and 240 control students (there are fewer observations for the state writing and math scores, the pre-composite *Reading Edge* score, and the pre-SFA assessment).

* Based on regression of characteristic on left on whether the student was randomly selected for FFW conditional on the student's randomization pool.

Table 3:
Numbers of Days Trained on FFW and Percent Who Have Reached Completion Criteria, Among Those Who Actually Trained on FFW by Flight

	Flight 1	Flight 2
Total Potential FFW Training Days	37.41	30.49
	[0.49]	[1.11]
FFW Language /Middle School		
Number of Days Trained	24.82	25.74
	[6.99]	[6.54]
Number of Complete Days	20.03	19.89
	[9.04]	[8.48]
FFW Language-to-Reading		
Number of Days Trained	11.99	9.42
	[4.47]	[3.15]
Number of Complete Days	8.40	8.42
	[4.31]	[3.38]
Total		
Total Days Trained	34.93	28.00
	[6.91]	[5.69]
Total Complete Days	27.11	21.91
	[10.36]	[8.83]
Proportion of Students With 20+ Complete Days	0.76	0.67
	[0.43]	[0.47]
Proportion of Students with 30+ Complete Days	0.58	0.27
	[0.49]	[0.45]
Proportion of Students with 80%+ Completion on a Majority of Exercises	0.53	0.42
	[0.50]	[0.49]
Proportion of Students with 20+ Complete Days and 80%+ Completion on a Majority of Exercises	0.51	0.38
	[0.50]	[0.49]
Proportion of Students with 30+ Complete Days and 80%+ Completion on a Majority of Exercises	0.44	0.15
	[0.50]	[0.35]
Maximum Number of Observations	140	129

Notes: Standard deviations in brackets. There were 8 students selected for FFW who never trained; they are not included in this table; 4 students who were not selected for FFW but nonetheless trained are included in the table. In the first flight 140 students trained on FFW Language, and 118 students trained on FFW Language-to-Reading. In the second flight 129 students trained on FFW Language or Middle School, and 31 students trained on FFW Language-to-Reading.

Table 4
Mean Pre- and Post-Test Scores for Treatments and Controls with Standard Errors in Parentheses *Reading Edge*, CELF-3-RP, *Success for All* (SFA), and State Reading Assessments

	Composite *Reading Edge* Test Scores		
	Pre-Scores	Post-Scores	Pre-Post Difference
FFW Students	52.54	73.97	21.43
	(1.93)	(1.61)	(1.54)
Controls	54.87	72.53	17.67
	(1.97)	(1.75)	(1.74)
Treatment-Control Difference	-2.33	1.44	3.76
	(2.76)	(2.37)	(2.31)
	Overall CELF-3-RP Scores (NCE)		
	Pre-Scores	Post-Scores	Pre-Post Difference
FFW Students	25.74	32.09	6.35
	(2.86)	(2.81)	(2.29)
Controls	25.43	31.01	5.59
	(2.75)	(2.53)	(2.24)
Treatment-Control Difference	0.31	1.07	0.76
	(3.97)	(3.78)	(3.20)
	SFA Assessment		
	Pre-Scores	Post-Scores	Pre-Post Difference
FFW Students	3.84	4.11	0.27
	(0.10)	(0.10)	(0.03)
Controls	3.78	4.03	0.25
	(0.10)	(0.10)	(0.03)
Treatment-Control Difference	0.06	0.08	0.02
	(0.14)	(0.14)	(0.04)
	State Standardized Reading Test (Percentile)		
	Pre-Scores	Post-Scores	Pre-Post Difference

FFW Students	38.81	44.57	5.75
	(1.31)	(1.61)	(1.48)
Controls	38.63	43.03	4.39
	(1.29)	(1.63)	(1.39)
Treatment-Control Difference	0.18	1.54	1.36
	(1.85)	(2.29)	(2.04)

Notes: Standard errors in parentheses. There are 244 treatments and 219 controls in the *Reading Edge* sample; 43 treatments and 43 controls in the CELF-3-RP sample, 197 treatments and 176 controls in the SFA sample; and 237 treatments and 217 controls in the State Test sample. Note that all samples only include students not missing either the pre- or post-test in question (and the *Reading Edge* sample only includes those students not missing any components of the pre-test).

Table 5a:
Regression Estimates of the Intent-to-Treat Effect with FFW for *Reading Edge* and CELF-3-RP Test Scores

	Reading Edge			CELF-3-RP		
	(1)	(2)	(3)	(4)	(5)	(6)
Selected for FFW	1.807	3.276	3.091	0.998	0.970	0.693
	(2.272)	(1.884)	(1.873)	(3.506)	(2.825)	(2.934)
CELF-3-RP Pre-Test					0.607	0.601
					(0.088)	(0.091)
Reading Edge Pre-Test		0.499	0.489			
		(0.034)	(0.034)			
Female			-5.256			-0.617
			(1.928)			(2.992)
African American			5.748			9.266
			(3.943)			(10.457)
Hispanic			1.733			6.577
			(3.742)			(10.386)
R^2	0.090	0.384	0.398	0.118	0.443	0.459
Number of Observations	485	485	485	89	89	89

Notes: Standard errors are in parentheses. The dependent variable in columns (1)-(3) is the composite of the post-FFW *Reading Edge* test score; the dependent variable in columns (4)-(6) is the normal curve equivalent of the post-FFW CELF-3-RP composite score. All specifications include a constant and controls for the student's randomization pool. Columns (2) and (3) also include dummy variables indicating which components of the pre-FFW composite *Reading Edge* score were missing. Columns (5) and (6) also include a dummy variable indicating if the CELF-3-RP pre-test is missing and column (6) also includes two evaluator dummy variables.

Table 5b:
Regression Estimates of the Intent-to-Treat with FFW for SFA and State Standardized Reading Assessments

	SFA			State Reading Assessment		
	(1)	(2)	(3)	(4)	(5)	(6)
Selected for FFW	0.070	0.032	0.031	1.728	1.538	1.146
	(0.103)	(0.038)	(0.038)	(2.157)	(1.847)	(1.833)
SFA Pre-Test		0.920	0.919			
		(0.019)	(0.019)			
State Reading Pre-Test					0.658	0.653
					(0.052)	(0.051)
Female			0.012			1.349
			(0.039)			(1.879)
African American			0.018			-10.016
			(0.082)			(3.862)
Hispanic			-0.072			-13.038
			(0.081)			(3.647)
R^2	0.489	0.930	0.931	0.147	0.376	0.395
Number of Observations	374	374	374	454	454	454

Notes: Standard errors are in parentheses. All specifications include a constant and the student's randomization pool. The dependent variable in columns (1)-(3) is the March assessment for the 3rd and 5th graders and the June assessment for the 4th and 6th graders (i.e., the assessment immediately following the end of the relevant FFW flight). The SFA Pre-test is the January assessment for the 3rd and 5th graders and the March assessment for the 4th and 6th graders. The dependent variable in columns (4)-(6) is the state standardized reading test (district percentile score) for the 2002-2003 school year. The state reading pre-test is from the 2001-2002 school year.

Table 6:
Instrumental Variables (IV) Estimates of the Effect of Treatment-on-the-Treated with FFW for *Reading Edge*, CELF-3-RP, SFA, and State Standardized Reading Test Scores

	Definition of "Treatment"		
	Number of Complete Days of Training	20+ Days Training & 80%+ on Exercises	30+ Days of Training & 80%+ on Exercises
Outcome	(1)	(2)	(3)
Reading Edge Composite Score	0.126	7.006	10.548
	(0.076)	(4.201)	(6.349)
CELF-3-RP Composite Score	0.028	1.720	3.654
	(0.118)	(7.281)	(15.532)
SFA Assessments	0.001	0.076	0.125
	(0.002)	(0.094)	(0.155)
State Standardized Reading Percentile Score	0.052	2.365	4.345
	(0.079)	(4.340)	(6.633)

Notes: Standard errors are in parentheses. The specifications are the same as those in columns (3) and (6) of Tables 5a and 5b. The definitions of treatment in columns (2) and (3) represent those who trained at least 20 (col. (2)) or 30 days (col. (3)) and completed 80% or more on a majority of exercises, including at least one sound exercise and at least one word exercise. There are 485 observations in the rows with *Reading Edge* outcomes, 89 observations in the rows with CELF-3-RP outcomes, 374 observations in the row with the SFA assessments, and 454 observations in the row with the State Standardized Reading Percentile Score.

Appendix Table 1:
Statistical Profile of an Average Regular Elementary School in the District, an Average Elementary School in the *Success-for-All* (SFA) Sample, and the Four Schools in FFW Evaluation, 2000-2001 School Year

	Average Elementary School in District	Average Elementary School in SFA Sample	Schools in Evaluation			
			A	B	C	D
Total Enrollment	625	635	457	795	482	996
% African American	39%	42%	54%	53%	6%	25%
% Hispanic	55%	51%	35%	42%	94%	59%
% Free- or Reduced Lunch Eligible	76%	73%	52%	79%	45%	61%
% K-12 Students with Non-English Home Language	60%	56%	39%	44%	99%	62%

Note: There are 26 "regular elementary schools" in the first column; the averages are weighted by the school's total enrollment. The average elementary school in the district averages exclude 7 charter and alternative schools. The average elementary school in the SFA sample excludes 9 elementary schools for which we could not obtain reliable SFA data as well as the schools that do not administer SFA. Note that the letter designations for the schools do not necessarily match those in Appendix Table 3.

Source: Authors' calculations from data provided by the State and the State's website.

Appendix Table 2
Reasons for Student Non-Participation in FFW Evaluation

	Flight 1	Flight 2
Total Eligible	389	371
Reason for Exclusion from Evaluation		
No Parental Consent	55	32
Behavioral Issue	45	44
Wildcard In FFW	1	1
Wildcard Out of FFW	1	0
Already Had Transferred Out of School	18	36
Family on Long Trip	1	0
Already Completed FFW	0	14
Total in Evaluation	268	244

Appendix Table 3:
Characteristics of Students by Participation in FFW Evaluation

	In FFW Evaluation	Not in FFW Evaluation	p-value of difference
Female	0.516 [0.500]	0.427 [0.496]	0.023
African American	0.267 [0.443]	0.347 [0.477]	0.025
Hispanic	0.654 [0.476]	0.556 [0.498]	0.009
3rd Grade	0.289 [0.454]	0.270 [0.445]	0.588
4th Grade	0.281 [0.450]	0.323 [0.468]	0.242
5th Grade	0.234 [0.424]	0.218 [0.413]	0.609
6th Grade	0.195 [0.397]	0.189 [0.393]	0.850
School A	0.264 [0.441]	0.153 [0.361]	0.001
School B	0.242 [0.429]	0.343 [0.475]	0.004
School C	0.303 [0.460]	0.371 [0.484]	0.060
School D	0.191 [0.393]	0.133 [0.340]	0.046
Composite State Reading Score	38.08 [19.12]	38.68 [20.08]	0.788
Composite State Writing Score	43.37 [23.92]	47.14 [24.11]	0.059

Total Math Score	46.07	39.18	0.002
	[26.52]	[26.32]	
Number of Observations	512	248	

Notes: Standard deviations in brackets. There are 460 and 210 observations for the writing scores and 466 and 193 observations for the math test scores for the participants and non-participants respectively. Note that the letter designations for the schools do not necessarily match those in Appendix Table 1.

Appendix Table 4:
Percentage Complete for Each FFW Exercise and
Proportion of Students Who Completed At Least 80% for Each Exercise

	Average Percent Completed	Percent Students Completing 80%+
FFW Language		
Circus Sequence	55.8 [34.0]	31.9 [46.7]
Phoneme Identification	62.1 [28.9]	31.1 [46.4]
Old MacDonald's Flying Farm	51.4 [34.3]	29.9 [45.9]
Phonemic Word	65.5 [40.0]	56.8 [49.6]
Phonemic Match	60.9 [34.5]	45.2 [49.9]
Block Commander	77.1 [26.3]	53.5 [50.0]
Language Comprehension Builder	80.3 [35.0]	78.8 [40.9]
FFW Middle School		
Sweeps	43.5 [36.4]	26.8 [44.7]
IDs	47.0 [27.0]	19.6 [40.1]
Streams	77.1 [27.2]	67.9 [47.1]
Matches	76.6 [32.5]	71.4 [45.6]
Cards	89.3 [21.4]	85.7 [35.3]
Stories	46.7 [29.0]	16.1 [37.1]
FFW Language-to-Reading		
Trog Walkers	39.5 [25.6]	6.7 [25.0]
Treasures in the Tomb	35.6 [17.8]	1.3 [11.5]
Polar Cop	14.8 [8.0]]	0.0 [0.0]
Bug Out	70.9 [26.5]	48.0 [50.1]
Start-Up Stories	67.1 [29.5]	45.3 [49.9]

Notes: Standard deviations in brackets. There are 241 observations for the FFW Language exercises; 56 for the FFW Middle School exercises; and 150 for the FFW Language-to-Reading exercises. The table represents students from both flights.