

Evaluation of the Voyager Universal Literacy System[®]

By



Prepared for:

The Council of the Great City Schools, Washington, D.C.
and Voyager Expanded Learning, Dallas, Texas

September 2003

Authors:

Dr. Joy Frechtling
Gary Silverstein
Dr. Xiaodong Zhang

TABLE OF CONTENTS

Chapter	Page
ABSTRACT5
1 INTRODUCTION6
The Voyager Universal Literacy System®6
Study Purpose and Methodology9
Site Selection9
Assessment of Early Reading Skills12
Study Activities14
Structure of the Report15
2 INSTRUCTIONAL CHARACTERISTICS OF VOYAGER AND NON-VOYAGER SCHOOLS16
Comparability With Regard to School and Classroom Environment16
Overall Classroom Environment16
Presence of Other Educational Reforms17
Parental Involvement in School-Related Activities18
Students' Pre-Kindergarten Exposure to Reading19
Principals' Assessments of Their Schools' Strengths and Challenges19
Comparability With Regard to Reading Instruction20
Students' Mastery of Literacy Skills21
Amount of Time Devoted to the Morning Reading Block21
Use of Small Group Instruction22
Principal Involvement in the Kindergarten Reading Program22
3 IMPACT OF VOYAGER ON STUDENT ACHIEVEMENT23
Methodology23
Sample23
Design25
Analysis25
Findings28
4 CONCLUSIONS39
REFERENCES40

TABLE OF CONTENTS—CONTINUED

LIST OF APPENDICES

Appendix		Page
A	Detailed Regression Results for Program Effects	41
B	Detailed Regression Results for Implementation Effects	43

LIST OF TABLES

Table		Page
1-1	Characteristics of Voyager and non-Voyager schools in Cleveland	11
1-2	Characteristics of Voyager and non-Voyager schools in Washington, DC	12
3-1	Comparison of classroom characteristics of Voyager and non-Voyager samples	24
3-2	Comparison of attrition rates between students in treatment and comparison schools	24
3-3	Distribution of implementation scores for teachers	28
3-4	Descriptive statistics in pre-test student achievement level (raw scores)..	29
3-5	Descriptive statistics in post-test student achievement level (raw scores)	29
3-6	Pre-test student CTOPP achievement mean in raw, percentile, and standard scores	30
3-7	Post-test student CTOPP achievement mean in raw, percentile, and standard scores	30
3-8	Student progress in Voyager schools	31
3-9	Comparison of progress between students in treatment and comparison schools (overall)	32

TABLE OF CONTENTS—CONTINUED

LIST OF TABLES—CONTINUED

Table		Page
3–10	Comparison of progress between students in treatment and comparison schools at district level: District 132
3–11	Comparison of progress between students in treatment and comparison schools at district level: District 233
3–12	Comparison of progress between students in treatment and comparison schools at school level: District 1, school pair 133
3–13	Comparison of progress between students in treatment and comparison schools at school level: District 1, school pair 234
3–14	Comparison of progress between students in treatment and comparison schools at school level: District 2, school pair 134
3–15	Comparison of progress between students in treatment and comparison schools at school level: District 2, school pair 234
3–16	Regression summary of the program effect models (N=398)35
3–17	Regression summary of the program implementation models (raw scores, N=202))36
3–18	Regression summary of the program implementation models (grouping, N=202)37
3–19	Pre and post raw scores by implementation level37
3–20	Correlations between evaluation test battery and the VIP38

LIST OF EXHIBITS

Exhibit

1–1	Form used for conducting observations8
-----	---	----

ABSTRACT

The evaluation of the Voyager Universal Literacy System[®] (Voyager) was designed to provide a rigorous assessment of the effectiveness of the program with beginning readers. Using a quasi-experimental design, researchers conducted a systematic evaluation of changes in kindergarten students' command of early reading skills in four Voyager and four comparison schools in two districts. The study provides striking evidence of the efficacy of the Voyager program. Overall and for three out of the four pairs of schools examined, a significant difference was found in favor of the Voyager classrooms. Effect sizes of the program ranged from 0.3 to 1.32 in seven test instruments, with an average of 0.62. The study also found that fidelity of implementation was an important variable. The greater the fidelity of the implemented program to the Voyager plan, the greater the difference in literacy scores. No consistent differences were found relating to student or teacher characteristics.

1. INTRODUCTION

The No Child Left Behind legislation requires that all schools take proactive steps to assure that all of their students are reading at grade level, and schools—especially those with large at-risk populations—are seeking ways to do so. At the request of Voyager Expanded Learning and the Council of the Great City Schools, Westat has conducted an evaluation of the ability of the Voyager Universal Literacy System® (Voyager) to help schools meet this goal. The evaluation was designed to determine the effectiveness of the program for kindergarten students in urban schools with at-risk populations. Kindergarten was targeted because it is when teachers must begin to deal with the background inadequacies that their young students bring to the challenge of learning, that is, the beginning of the achievement gap. With the use of a quasi-experimental design, this evaluation examined the comparative effects of Voyager on the learning of early literacy skills in kindergarten children.

This chapter provides background on the Voyager approach, describes the purpose of the study, and outlines the methodology used to examine the implementation and impact of the Voyager program.

The Voyager Universal Literacy System

Voyager is a comprehensive K–3 reading program designed to provide all students with the skills and knowledge needed to become proficient readers. This system includes:

- A core K–3 reading curriculum;
- A progress monitoring system that measures each student’s reading progress and identifies struggling readers;
- A struggling reader intervention that includes additional instruction time and summer school;
- Professional development for teachers, principals, and specialists;
- A home study curriculum;
- Incentives for educators; and
- Technology enhancement activities that provide practice on phonological skills and enrichment in comprehension, fluency, and writing.

The Voyager curriculum provides a detailed scope and sequence of skill development. It was designed to be aligned with the critical elements of reading instruction identified in *No Child Left Behind*: phonemic awareness, phonics, fluency, vocabulary, and comprehension. At the kindergarten level, the curriculum is divided into daily instructional blocks of 2 hours. The suggested daily schedule includes the Friendship Circle (whole group instruction; 20 minutes) Reading Stations (differentiated, small group instruction including two student-led independent stations and one teacher-led station; 70 minutes), and Writing Connection (teacher facilitated writing activities; 30 minutes).

Voyager uses a progress-monitoring system to assess each student on indicators of early literacy. This system includes Voyager Indicators of Progress™ (VIP) and Assessment Checkpoints. VIP is a timed oral measure that teachers administer to students once in the fall, twice in the winter, and once in the spring. VIP scores help teachers evaluate individual students' progress and customize instruction for students at various levels of skill development. VIP scores determine the student groupings for Reading Stations. VIP is also an online data management and reporting system that generates reports on reading progress at the student and classroom level. Assessment Checkpoints, criterion-referenced skill assessments, are administered to students every 6 weeks. Information from the assessments help teachers adjust group composition and identify specific problems or skills that need to be retaught.

Voyager professional development includes an initial 2-day training session for district and campus coaches and a 3-day training session for teachers. There are also eight 3-hour professional development modules for ongoing development throughout the school year. The initial training focuses on classroom management, administering the VIP, grouping students, delivering the curriculum, and Voyager reading instruction. The ongoing sessions expand on all of these areas: topics are introduced by videotape, and the campus coaches facilitate group processing and practice. Some modules include exercises such as reciprocal classroom observations. In addition, Voyager staff periodically observe teachers during the reading block to assess the extent to which the Voyager approach is “implemented with fidelity.” A copy of the form used to conduct these observations is provided in exhibit 1–1.

Exhibit 1–1.—Form used for conducting observations

INSTRUCTIONAL FIDELITY INDEX		UNIVERSAL LITERACY IMPLEMENTATION Instructional Fidelity Index (IFI)	
<p>Check One</p> <p><input type="checkbox"/> Self</p> <p><input type="checkbox"/> Observer</p> <p>Signature _____</p> <p>Time Spent Observing _____</p>		<p>Name _____ Grade _____ Date _____</p> <p>School _____ Visit Number _____</p> <p>Use the Instructional Fidelity Index to determine the level of implementation of each component of the Universal Literacy System and to note change over time. Circle the number that best represents your judgment of each component's use.</p> <ul style="list-style-type: none"> • Circle 0 for NOT IMPLEMENTED if that component of Universal Literacy is not in use. • Circle 1 for IMPLEMENTING WITH LIMITED FIDELITY if there is minimal evidence of the component. • Circle 2 for IMPLEMENTING WITH SOME FIDELITY if there is some evidence of the component. • Circle 3 for IMPLEMENTING EFFECTIVELY if there is evidence of considerable skill in the component. 	
<p>1. The teacher provides instruction in small same-ability groups in Teaching Station.</p> <ul style="list-style-type: none"> * Teacher provides instruction that is targeted to group's learning level. * Teacher teaches deliberately and persistently. • Teacher follows Teaching Station lesson from Curriculum Guide. (May add extension activities for On-Track students after basic lesson.) • Students respond individually and as a group. • Teacher monitors student progress and checks for mastery. • Teacher varies size of group and length of instruction depending on students' needs. 	<p>0 1 2 3</p>		
<p>2. The teacher uses Curriculum Guide consistently and appropriately.</p> <ul style="list-style-type: none"> * Teacher closely follows daily lesson plan provided in Guide. * Teacher uses instructional models (from the Skill Development and Teaching Station sections of the Guide), including correction procedures, with complete fidelity. • Teacher demonstrates familiarity with the day's lesson plan. • Teacher responds knowledgeably to questions about the day's priority skills and the learning of each student. 	<p>0 1 2 3</p>		
<p>3. The teacher has established independent Reading Stations and uses them appropriately.</p> <ul style="list-style-type: none"> * Students complete Reading Station activities. • Teachers prepare Team Leaders for Reading Stations. • Teachers debrief independent Reading Station activities with students to insure that learning has occurred. 	<p>0 1 2 3</p>		
<p>4. The teacher (or other campus professional) provides appropriate student intervention for Struggling and low Emerging Readers.</p> <ul style="list-style-type: none"> * Struggling and low Emerging Readers receive 10–30 minutes of additional instruction (Reading Intervention lesson). * Struggling and low Emerging Readers in Grades 1 and 2 receive Extended-Time Intervention. 	<p>0 1 2 3</p>		
<p>* Priority Items</p>			

Study Purpose and Methodology

Voyager Expanded Learning and the Council of Great City Schools contracted with Westat to conduct an evaluation of the Voyager approach in several school districts.¹ The study conducted during the 2002–03 school year addressed the following questions:

- What is the impact of the Voyager Universal Literacy System® on student literacy learning? Does reading performance in participating schools differ from that in comparable non-participating schools in the same district?
- Is the program equally effective for students from different backgrounds—gender, race, ethnicity, economic status, and English language skills?
- How does the level of implementation affect outcomes?
- What is the relationship between student performance on the evaluation test battery and VIP?

This section describes the methodology that was used to address these study questions. Specifically, it provides information about the process used to select study sites, assesses the comparability of the treatment and comparison schools, and describes steps taken to examine the implementation and impact of the Voyager program.

Site Selection

Westat worked with Voyager Expanded Learning and the Council of Great City Schools to identify urban school districts that would be interested in participating in the study. In order to participate, districts had to agree to:

- Identify schools (with comparable demographic and educational characteristics) that would be willing to begin using Voyager with their kindergarten students in fall 2002;

¹ Originally, Westat was asked to follow a cohort of students from kindergarten through third grade in three urban school districts. However, Voyager ultimately decided to scale back the focus of the study to an examination of its reading program at the kindergarten level in two urban school districts.

² To compensate them for their efforts, comparison schools received an annual stipend of \$500 to be used in a manner determined by the principal and participating faculty.

- Allow these schools to be randomly assigned to either a treatment group (that provided Voyager to all kindergarten students) or a comparison group (that would continue using the existing reading program with its kindergarten students);²
- Allow Westat to conduct annual site visits to both treatment and comparison schools;
- Provide existing school and student-level data for treatment and comparison schools; and
- Allow students in both treatment and comparison schools to be periodically tested on the evaluation’s reading assessment battery.

Both of the districts (Cleveland and Washington, DC) that were selected to participate in the study provided us with two lists of schools—one of potential treatment (Voyager) schools and one of potential comparison (non-Voyager) schools.³ For each school nominated, we asked the districts to supply us with the following information:

- Number of students,
- Number of teachers,
- Number of full-day and half-day kindergarten classrooms,
- Student gender (percent),
- Student race/ethnicity (percent),
- Students receiving free/reduced-price lunch (percent),
- Students who are limited English proficient (percent),
- Students with individualized education plans (IEPs) (percent),
- Student achievement on an existing local or national assessment, and
- Student mobility rate.

We attempted to select schools that were comparable on both school and student characteristics. Specifically, we wanted schools that had at least two, preferably three, kindergarten classrooms. As indicated in tables 1–1 (Cleveland) and 1–2 (Washington, DC), most of the selected schools had three full-day kindergarten classrooms with comparable percentages of male and female students and African American students (over 90 percent in all eight schools). We also matched schools by

³ We had initially hoped to have one list of potential schools from which random assignment could be carried out. For a variety of reasons, the school districts were unable to go along with this approach. We instead chose a quasi-experimental design by matching the treatment and comparison groups.

percentages of students with limited English proficiency, IEPs, and receipt of free or reduced-price lunch. One of the Cleveland schools had a high mobility rate (14.3), so we ensured that its comparison school had a similarly high rate (17.7). Finally, we selected schools that had similar levels of reading achievement on the district’s standardized achievement test. Cleveland scores indicate the “proficient rate” on the Ohio Proficiency Test, and Washington, DC, scores indicate the Mean Normed Curve Equivalent on the SAT-9. Achievement data were from the most recent test administration, i.e., the 2000 school year.

Overall, the matches for the Cleveland schools are very good. While adequate, the Washington sample had some shortcomings (see, for example, percentages of students receiving free and reduced-price lunch). In addition, we encountered some last minute problems in Washington when a school dropped out shortly before the school year started. The substitution school represents the best match that could be made under the circumstances.

Table 1–1.—Characteristics of Voyager and non-Voyager schools in Cleveland

Characteristic	School A (Voyager)	School B (Comparison)	School C (Comparison)	School C (Comparison)
School Characteristics				
Total number of students	499	565	273	439
Total number of classroom teachers	29	45	21	32
Number of full-day kindergarten classrooms	5	3	3	3
Number of part-day kindergarten classrooms	0	0	0	0
Student Characteristics				
Percent female	51	49	53	50
Percent male	49	51	47	50
Percent American Indian/Alaska Native	0	0	0	0
Percent Asian/Asian American	0	0	0	0
Percent Black/African American	100	100	100	97
Percent Native Hawaiian or Pacific Islander	0	0	0	0
Percent White	0	0	0	3
Percent Hispanic	0	0	0	0
Percent receiving free/reduced-price lunch	93	86	94	89
Percent limited English proficient	0	0	0	0
Percent with IEPs	12	10	8	8
Reading scores*	22	25	34	40
Mobility rate	14	18	1	1

*Ohio Department of Education Interactive Local Report Card, 2000–2001 school year.

NOTE: Percents have been rounded.

Table 1–2—Characteristics of Voyager and non-Voyager schools in Washington, DC

Characteristic	School A (Voyager)	School B (Comparison)	School C (Comparison)	School C (Comparison)
School Characteristics				
Total number of students	375	371	356	484
Total number of classroom teachers	19	19	30	24
Number of full-day kindergarten classrooms	3	3	2	3
Number of part-day kindergarten classrooms	0	0	0	0
Student Characteristics				
Percent female	51	54	46	48
Percent male	49	46	54	52
Percent American Indian/Alaska Native	0	0	0	0
Percent Asian/Asian American	0	1	1	<1
Percent Black/African American	100	95	95	91
Percent Native Hawaiian or Pacific Islander	0	0	0	0
Percent White	0	0	1	<1
Percent Hispanic	0	3	4	8
Percent receiving free/reduced-price lunch	71	90	77	63
Percent limited English proficient	0	1	3	3
Percent with IEPs	16	6	11	9
Reading scores*	54	46	51	58
Mobility rate	NA	NA	NA	NA

*DC Public Schools website, 2001–02 school year.

NA = not available.

NOTE: Percents have been rounded.

Assessment of Early Reading Skills

Westat, in consultation with the Voyager Advisory Team,⁴ selected a test battery to assess letter identification, phonological awareness, and emerging reading skills. All three tests demonstrate good psychometric properties, are appropriate for young children, and are short to administer. On average, each child spent 30 minutes on the test battery. The assessment consisted of the following subtests.

- **Dynamic Indicators of Basic Early Literacy Skills™ (DIBELS)—Letter Naming Fluency subtest.** The DIBELS (Good and Kaminsky, 2002) tests are a set of standardized, individually administered measures of early literacy development. The Letter Naming

⁴ Members of the Advisory Team are Dr. Roland Good, Dr. Edward Kame’enui, Dr. Deborah Simmons, and Dr. Sharon Vaughn.

Fluency subtest is a page of upper and lower case letters arranged in a random order. The assessor asks the student to name as many letters as he/she can. Students are told if they do not know a letter, they will be told it. The assessor gives the student 1 minute to produce as many letter names as he/she can. According to the test benchmarks, students at the end of their kindergarten year should be able to name more than 40 letters in 1 minute to be considered “at low risk” for difficulty in achieving early literacy goals. Students who can name between 29 and 40 are “at some risk,” while those who can name fewer than 29 letters in 1 minute are “at risk.”

- **Comprehensive Test of Phonological Processing (CTOPP)—Elision, Blending Words, Blending Nonwords, and Segmenting Words subtests.** The CTOPP (Wagner, Torgesen, and Bryant, 1999) assesses phonological awareness, phonological memory, and rapid naming skills. The Voyager assessment battery used four subtests of CTOPP: Elision, Blending Words, Blending Nonwords, and Segmenting Words. The Elision test requires segmenting spoken words into smaller parts, a skill that is a precursor to identifying how sounds map onto letters in words. This 20-item test measures the extent to which an individual can say a word, then say what is left after dropping out designated sounds. The 20-item Blending Words test measures an individual’s ability to combine sounds to form words. The student listens to audiocassette-recorded sounds and is asked to combine the sounds together to make a whole word. The 18-item Blending Nonwords test requires combining speech sounds (again, heard on an audiocassette) into “made-up” words. This test is less affected by an individual’s vocabulary, and has higher memory demands, than the Blending Words test. The 20-item Segmenting Words subtest measures an individual’s ability to say the separate phonemes that make up a word. The student is asked to repeat a word (spoken by the assessor) and then say it one sound at a time.
- **Woodcock Reading Mastery Test Revised (WRMT-R)—Word Identification and Word Attack subtests.** The WRMT-R measures basic decoding skills. The Word Identification test requires the student to identify isolated words that appear in large type in the test book. As subjects proceed through the items, they encounter words that appear less and less frequently in written English. For an answer to be scored correct, the subject must produce a natural reading of the word within about 5 seconds. The subjects do not need to know the meaning of the words. The Word Attack test requires the subject to read either nonsense words or words with very low frequency in the English language. The test measures the subject’s ability to apply phonic and structural analysis skills in order to pronounce words with which he or she is unfamiliar.

These tests were selected because of their technical adequacy and connection with the most important constructs in kindergarten. The constructs tapped are highly predictive of later reading achievement as well as the factors most closely associated with effective reading instruction in kindergarten. These constructs represent the foundation areas of reading instruction in kindergarten that can be reliably measured: letter knowledge, letter naming fluency, phonological processing, and knowledge of word reading and word attack. As students progress through the grades, other critical elements of reading instruction can and should be measured. These include fluency in connected text and reading comprehension. Kindergarten students are not expected to be reading fluent text with automaticity and comprehending this text, and thus these elements were not tested during kindergarten. However, we intend to measure both fluency and comprehension in first grade and in later grades when students are expected to perform these elements of reading, and when they can be reliably measured.

Study Activities

Teachers in the treatment group received training and materials from Voyager in August 2002, and began implementing the program with their students in early September 2002.⁵ Testing on students' pre-intervention literacy abilities was carried out by Westat in October 2002 in both Cleveland and Washington. Most students were tested within a 2-week period, with a few make-ups extending beyond that time. We tested all available students, since no school met or exceeded a 75-pupil cap put on the assessment sample. Prior to the testing, experienced assessors received approximately 3 days of training from Westat; the first 2 days focused on developing an understanding of the assessments and the last, on practicing with young children. Performance of the assessors was audited by Westat's senior trainer on a sampling basis. Results of these tests indicate that at the outset of the school year, there were no differences either overall or within the districts between the treatment and comparison groups. In May 2003, Westat re-administered the common core reading battery to those students who had been tested at the beginning of the study.

In spring 2003, Westat conducted a site visit to each of the eight schools participating in the study to observe the classrooms and interview teachers and principals. In Voyager schools, the site visitors also focused on the level of implementation of each component of the Voyager system (e.g., the extent to which the teacher followed the Curriculum Guide, and used the Reading and Teaching Stations appropriately). In comparison schools, the site visitors determined which curriculum was being used and the extent to which the teachers followed that curriculum. While these visits were

⁵ Kindergarten teachers at one school in Washington, DC, did not begin implementing Voyager until October 2002.

too limited to provide a detailed examination of fidelity of the implementation of the Voyager program, measures of program implementation were provided to Westat by Voyager staff before the evaluation results were known using data collected using the IFI form (see exhibit 1–1). These site visits produced valuable qualitative data regarding school characteristics, reading program characteristics, teacher qualifications and experience, and other factors that could influence student reading outcomes.

Structure of the Report

The remainder of this report summarizes findings from the site visits and analyses of the student assessment data. Chapter 2 compares the academic characteristics of the Voyager and non-Voyager schools—with special emphasis on similarities and differences in the educational environment of kindergarten students—and provides information on the implementation experiences of the Voyager approach in the four treatment schools. Chapter 3 examines the impact of Voyager on student achievement. Chapter 4 provides a conclusion summary.

2. INSTRUCTIONAL CHARACTERISTICS OF VOYAGER AND NON-VOYAGER SCHOOLS

A primary purpose of this study is to assess whether the use of Voyager results in improved student performance when compared with other kindergarten literacy programs. This assessment requires that the treatment and comparison schools be as similar as possible. As discussed in Chapter 1, the treatment and comparison schools were similar in terms of demographic characteristics and the pre-intervention literacy aptitudes of kindergarten students. This chapter documents the similarities and differences in the Voyager and non-Voyager schools with respect to their overall instructional environments. In so doing, it also provides some contextual information about the academic environment within which this study was conducted. Finally, this chapter provides information on the extent to which Voyager was implemented as intended in the four schools in the study.

Comparability With Regard to School and Classroom Environment

In an effort to understand the environment within which reading instruction was being provided in Voyager and non-Voyager classrooms, we conducted 1-day site visits to each school in the study. During these visits, which occurred in spring 2003, we interviewed the principal and kindergarten teachers—and observed a sample of kindergarten reading blocks—to assess the extent to which other environmental factors might explain differences in the end-of-year reading skills of the kindergarten students who participated in the study. Although these visits were not sufficient to conduct an exhaustive comparison of treatment and comparison instructional practices, they did provide an opportunity to make some general assessments about the similarities and differences of Voyager and non-Voyager schools. As is discussed throughout this section, the findings from these visits do not suggest that there were systematic or significant differences among the eight schools in the study.

Overall Classroom Environment

Without exception, the classrooms we visited were colorful and offered inviting learning environments. Each of the classrooms was well decorated with students' artwork and other schoolwork. (In one non-Voyager classroom, students' work was posted on the walls and the ceiling.) All of the classrooms had a good selection of educational manipulatives, picture and story

books, and toys. While some classrooms had one or two instructional aides, others were only staffed by the kindergarten teacher. The number of classroom computers varied in both treatment and comparison schools—e.g., two Voyager classrooms had three and four computers, respectively, while one non-Voyager classroom had five computers. Computers in both treatment and comparison schools had reading software.

Presence of Other Educational Reforms

Principals and teachers were asked to describe any educational reforms (other than Voyager), partnerships, and special education programs that might affect the reading abilities of kindergarten students. The purpose was to assess whether the presence of other instructional practices or reforms might explain any observed differences between Voyager and non-Voyager classrooms.

While none of the schools were involved in a comprehensive educational reform initiative, all described additional efforts that were underway to promote literacy and maximize the instructional opportunities made available to their students. For example, several Voyager and non-Voyager schools conducted after-school programs for students in need of remedial assistance (although this assistance was often not made available to kindergarten students), brought in volunteers (e.g., from local high schools, colleges, law firms, retirement communities) to tutor kindergarten students, and established partnerships with visible community entities (e.g., local television stations, a well-known basketball coach at a local university). Staff in several of the schools had written grants to secure additional supplies or fund special projects. In addition:

- In Cleveland, both of the comparison schools indicated that they were part of an effort to institute 90-minute literacy blocks. (This issue was not raised in the Voyager schools, since the program requires that kindergarten teachers devote 2 hours per day to Voyager activities.)
- Teachers in one of Cleveland’s comparison schools were taking steps to delineate beginning and end-of-year skills that could be used to align curriculum and expectations within and across grade levels.
- One of Cleveland’s comparison schools had established a partnership with Case Western University that allowed students at all grade levels who surpassed a reading goal (i.e., number of books read) to attend a basketball game.

- One of the Washington Voyager schools adopted a range of incentives to promote reading at all grade levels—including providing coupons for a free pizza to students who surpass their reading goals, and a hot air balloon ride to the student who reads the most books over the course of the school year.

The prominence of these practices is noteworthy, because they demonstrate that several of the treatment and comparison schools were taking additional steps to emphasize literacy and promote reading among their students. However, it appears that these practices were evenly distributed across the Voyager and non-Voyager schools. It also appears that there were no differences across Voyager and non-Voyager schools with respect to such educational factors as teachers' average tenure, teacher participation in non-Voyager professional development, and the amount of time allocated for lesson planning.

Parental Involvement in School-Related Activities

The extent to which parents are involved in their children's education can be an important predictor of how well those children perform in school. At the kindergarten level, such involvement can include taking the time to read to their children, assist with homework, volunteer in the classroom, and attend parent-teacher conferences. As such, significant differences in parental involvement across schools could serve to explain differences in the reading abilities of kindergarten students.

Most of the teachers that we interviewed indicated that lack of parental involvement was a serious problem for at least some of the students in their classroom. In both Cleveland and Washington, this problem was attributed to several factors—e.g., lack of time because parents were holding down a job, attending school, or looking after younger children; inability to read, especially for parents for whom English is not a primary language; and lack of interest in their child's education. However, many teachers also indicated that at least some of their students had parents who were involved in their children's schoolwork. While teachers' assessment of parental involvement varied within and across the eight schools in the study, these differences appeared to be evenly divided across the Voyager and non-Voyager schools. However, the extent of parental involvement in two comparison schools is worth noting.

- One of Cleveland's comparison schools had a significant number of students who resided in either a homeless shelter or a domestic violence center. The principal also identified a range of associated challenges—including parents waiting until April to enroll their children in

kindergarten, a high mobility rate (with one class in the school losing 15 of its 17 original students between September 2002 and April 2003), and a significant number of students being raised by foster parents or grandparents. As a result, lack of parental involvement in this school represented a significant barrier.

- One of the comparison schools in Washington, DC, attracted a significant number of students from outside the school’s boundary during the period of the study.⁶ According to the principal, this occurs because the school has a reputation for academic excellence. One of the kindergarten teachers estimated that 6 of her 16 students live outside the school’s boundary—and that the parents of these students tend to reinforce academic skills at home and volunteer for class field trips. This suggests a higher level of parental involvement and interest than the other schools in the study, since parents actively sought to enroll their children in this school for academic reasons.

Students’ Pre-Kindergarten Exposure to Reading

Teachers were asked to assess the extent to which their students had some exposure to reading prior to beginning kindergarten. Once again, results were mixed within and across Voyager and non-Voyager classrooms. Most teachers indicated that one-third to one-half of their students had been in a preschool environment (e.g., Head Start) that included a literacy component. While these students were not reading at the beginning of the school year, they were “pre-readers” who were familiar with books—e.g., they knew how to turn pages and could listen to a story. However, most teachers in Voyager and non-Voyager classrooms also indicated that a significant proportion of their students had never been read to prior to kindergarten and were unfamiliar with the written word. For example, teachers in a treatment school emphasized that some of their students began the school year without knowing their formal names (they were only aware of their nicknames) or their date of birth. In addition, a teacher in a comparison school indicated that at the beginning of the school year, the majority of her students could not differentiate between letters and numbers.

Principals’ Assessments of Their Schools’ Strengths and Challenges

Principals were asked to identify the greatest strengths and challenges associated with their

⁶ A lottery is used to determine which out-of-boundary students will attend the school. However, parents from outside the school’s boundary do not have to promise any special behaviors (e.g., monitoring their children’s homework) to get into the school.

elementary schools. Once again, the purpose was to gain an understanding of the similarities and differences in the environments in the schools in which the study was conducted. All but one of the principals cited the expertise and dedication of their instructional staff as a primary strength. For example, one principal (in a non-Voyager school) indicated that her kindergarten and first grade teams were “superior” and tended to collaborate with one another to “improve their art of teaching.” Another principal (in a Voyager school) described the willingness of her staff to participate in something new. Voyager and non-Voyager principals also cited many of the same challenges facing their elementary school—including a lack of parental support, a range of problematical student behaviors (e.g., learning disabilities, aggressive behavior), a lack of funding to hire content specialists, and crime and poverty in the surrounding neighborhood.

Comparability With Regard to Reading Instruction

The limited amount of time devoted to visiting individual schools makes it difficult to systematically compare the reading programs in the Voyager and non-Voyager programs.⁷ However, some observations regarding the treatment and comparison schools are worth noting.

- The instruction that we observed in each of the non-Voyager schools included activities that explicitly and directly addressed phonemic awareness, phonics, and sight words.
- All Voyager and non-Voyager teachers indicated that they read to students (during and outside of the formal reading block), integrated literacy skills into other lessons (e.g., science, social studies) throughout the day, and provided students with opportunities to “read” on their own.
- In most of the classrooms we observed, teachers assigned homework and expected it to be completed on time. In addition, several teachers in comparison schools were conducting weekly spelling tests.

⁷ It is worth noting another factor that complicates a comparison of the practices we observed in Voyager and non-Voyager schools—i.e., Voyager teachers were in the process of mastering a new reading curriculum, while the veteran teachers in non-Voyager schools already had a working knowledge of their reading program. In the comparison schools, the reading programs were familiar and well-established, having been in place a minimum of 3 years. As such, a Voyager teacher who appeared scripted in her first year of working with a new curriculum might come across more naturally in subsequent years.

- Although most Voyager and non-Voyager schools had a lead reading teacher, these specialists were generally not working with kindergarten students. In several schools, the lead reading teachers were concentrating their efforts on those students (generally 4th graders) who were about to participate in mandatory literacy assessments.

Other noteworthy similarities and differences in the reading programs offered across the classrooms we observed are summarized below.

Students' Mastery of Literacy Skills

Students seemed to vary in their mastery of the knowledge and skills of the reading program in both Voyager and non-Voyager schools. In all of the schools we visited, there would be a small group of students who had clearly mastered the content and knew what to do, a second group that went along with what appeared to be partial understanding of the lesson, and a third group that seemed disengaged from classroom activities (especially at independent learning centers). One non-Voyager teacher, for example, grouped the abilities of her 18 kindergarten students as follows:

- Five students needed supplemental assistance with basic letter sounds;
- Three students needed supplemental assistance on specific words;
- Eight students were borderline or on-track; and
- Two students knew all 40 sight words, could independently read text of stories, and exhibited good handwriting and comprehension skills.

Amount of Time Devoted to the Morning Reading Block

Teachers in the Voyager classrooms were devoting the required 90 minutes to a full reading block. While non-Voyager classrooms were spending less time on reading (i.e., 60 to 90 minutes), this comparison may not be valid because all of the treatment and comparison teachers we interviewed appeared to be integrating reading activities into other subjects on a daily basis. (Since only morning reading was observed in all the schools, it is difficult to comment on how much reading-related teaching takes place during other content lessons.) Most teachers described how

reading permeated all areas of their instruction. For example, one teacher in a comparison school described a recent science lesson on evaporation in which she devoted time to the spelling of the word “wet.”

Use of Small Group Instruction

Most of the teachers in the non-Voyager schools routinely used smaller groups to customize their literacy instruction. However, not all of these teachers used students’ aptitude to determine the composition of these groups. One non-Voyager teacher reported using small group instruction two to three times per week (while the rest of the class is working on an assignment) to work on reading and other academic subjects. She indicated that having students with mixed abilities in a small group setting allowed her to pay special attention to those remedial *and* advanced students in need of special assistance. Further, because she uses these small groups to tackle multiple subject areas, she found that some students who were good readers needed remedial assistance in mathematics (and vice versa). Two other non-Voyager teachers indicated that they deliberately organize their students in groups that include advanced, on-track, and struggling students—thereby providing an environment in which good readers can help the poor readers. During these small group activities, the teachers moved around the room working with individual students.⁸

Principal Involvement in the Kindergarten Reading Program

Most of the principals in Voyager and non-Voyager schools were not directly involved in the kindergarten reading program (other than to observe classes and provide support for the program). In Cleveland, the emphasis was clearly on assuring that 4th grade students were ready for their literacy assessments, and some principals were involved with those students (one actually provided tutoring to 4th graders). However, all of the principals supported reading at the kindergarten level by seeking additional resources for literacy (e.g., tutors or special programs) in their schools. In addition, several principals promoted reading at all grade levels by offering such special incentives as free pizzas and hot air balloon rides. (A sign posted in one non-Voyager kindergarten classroom proclaimed: “Meet your Accelerated Reader Goal for the month of May to pummel [name of principal] with a water balloon. The top reader gets an additional shot.”)

⁸ At least one Voyager teacher commented on the difficulty of working with the struggling readers group, since none of these students were in a position to model good reading skills.

3. IMPACT OF VOYAGER ON STUDENT ACHIEVEMENT

This chapter examines the impact of Voyager program on student achievement. It looks at early reading progress in treatment and comparison schools, assesses the relationship between implementation and progress, and examines the correlations between student performance on the assessment battery used in the evaluation and the Voyager Indicators of Progress (VIP), the curriculum-embedded assessment of reading skills. We found that:

- On average, students from Voyager schools scored higher than their counterparts in the comparison schools. The average effect sizes of the program ranged from moderate to large. Analyses at the school level also show that Voyager students gained more than their counterparts in three out of the four pairs.
- The fidelity of implementation of the Voyager program plays an important role in the observed outcomes. Students from high implementation classrooms scored higher than those in medium and low implementation classrooms.
- The program and its level of implementation have stronger effects on student achievement than do student and teacher characteristics.
- The assessment battery used in this evaluation has moderate but significant correlations with the VIP.

Methodology

This section details the procedures to address different methodological issues related to the study such as sampling, design, and analytical models.

Sample

Table 3–1 presents the means of classroom characteristics for the treatment and comparison samples.⁹ The results suggest no statistically significant difference between the Voyager and non-Voyager classrooms.

⁹ Tables 1-1 and 1-2 presented characteristics of Voyager and non-Voyager schools from the previous school year. The data in table 3-1 represent the descriptive statistics of the actual kindergarten treatment and comparison samples.

Table 3–1.—Comparison of classroom characteristics of Voyager and non-Voyager samples

Characteristic	Voyager classes (N=12)	Non-Voyager classes (N=12)	Difference	Significance (p value)
Class size	19.00	18.42	.58	.74
Teacher years of experience	7.25	5.87	1.38	.93
Teacher attendance rate	96.83	97.80	-.97	.28
Student attendance rate	93.35	96.20	-2.95	.12
Percent male	50.00	48.00	-1.92	.61
Percent black	98.00	96.00	1.75	.35
Percent students receiving free/reduced-priced lunch	84.00	93.00	-8.50	.25
Percent students limited English proficiency	.58	.92	-.34	.33
Percent students with individual education plan	4.41	1.50	2.91	.29

The study presents the results for a total of 398 students, 202 in the treatment group, and 196 in the comparison group who were both pre-tested in October 2002 and post-tested in May 2003. These 398 students represent 89 percent of the students initially pre-tested. Examination of school-level attrition rates shows that the attrition rates between treatment and comparison schools appeared to be comparable overall (table 3–2).

Table 3–2.—Comparison of attrition rates between students in treatment and comparison schools

District ID	School pair	Treatment	Comparison
1	1	6.9%	12.5%
1	2	8.0%	11.1%
2	3	13.7%	10.5%
2	4	12.3%	12.0%

The unit of analysis in this study is the student, with data coming from three sources. The student assessment data include a longitudinal sample of scores from pre-test and post-test for each individual student. Data on classroom characteristics were collected using teacher surveys in spring 2003. The implementation scores were reported by Voyager staff charged with monitoring, using classroom observations. Data collected at the classroom level were disaggregated to student level. Students in the same classroom share the same classroom-level variables.

Given the specific demographics of the sample, it is important to note that the findings can only be generalized to kindergarten students in inner-city schools with predominantly African American student bodies.

Design

As described in Chapter 1 of this report, the study can be characterized as a quasi-experimental design with pre-post test data. In other words, the Voyager schools were matched, to the best extent possible, with non-Voyager schools by various indicators. Each student was tested twice—once in October 2002 and once in May 2003. The difference in test scores between pre-test and post-test is the measure of program gain.

Analysis

The overarching purpose for this evaluation is to find out whether the Voyager program has any effect on student literacy learning. The Voyager organization was also interested in learning whether performance on the teacher-administered VIP correlated with performance on the externally administered evaluation assessment battery. In order to address these questions, we developed four research questions. This section details the approaches by which each question was addressed.

1. What is the impact of the Voyager Universal Literacy System on student literacy learning? Does reading performance in participating schools differ from that in comparable nonparticipating schools in the same district?

This first question can be further broken down into four subquestions.

1–1. What is the level of achievement for students in Voyager and non-Voyager schools in pre-test and post-test? While the primary focus of the study is about achievement gains, it is important to examine the level of student achievement in pre-test and post-test to gain an understanding of status.

1–2. What is the extent of gains for students in Voyager schools? The issue here is the difference of mean between the post-test and pre-test scores in the Voyager schools. Paired sample t-tests were performed on the mean difference of various assessments. Both statistical significance (p-value) and effect size were reported. The significance test provides a way of judging whether the difference is plausible and not due to chance. However, the statistical significance is dependent on sample size. When samples are very small, as is the case in this study, achieving statistical significance may be close to impossible. Therefore, we also used “effect size” as a measure of the difference.

1–3. Does student reading performance in participating schools differ from that in comparable nonparticipating schools at district and school level? This question was addressed by comparing the mean of the difference between treatment and comparison groups at the project, district, and school levels. Differences at the district level were compared using independent sample t-tests, whereas t-tests were not to be administered at school level due to the small number of observations within each school. Program effect sizes were then calculated.

1–4. To what extent can the gains be attributed to the program? Student learning can be affected by multiple factors, such as student socioeconomic status and teacher and school characteristics, other than the program. In order to control for these confounding factors, we developed a multiple regression ordinary least square (OLS) model to assess the net effect of the program.¹⁰

Model 1 (program effect)

$$\text{Gain score} = \alpha + \beta \text{ Voyager} + \delta_1 \text{Pctmale} + \delta_2 \text{Csize} + \delta_3 \text{Tchexp} + \delta \text{Pctfrl} + \varepsilon$$

In this model, the dependent variable is the difference in scores between pre-test and post-test for each student. The main independent variable is Voyager program, coded as “1” for students in the participating schools and “0” for those in nonparticipating schools. Percent male, class size, teacher’s experience teaching at K level, and percent of students receiving free and reduced-price lunch were introduced as control variables.¹¹ We are primarily interested in the parameter estimate for β , which measures the net program effect from the Voyager program. However, the parameter estimates in δ s are also interesting, because they represent influences from student and teacher variables within this particular population.

¹⁰ We explored the possibility of using a 3-level hierarchical model (level 1, time; level 2, student; level 3, class). The advantage of using the multilevel model lies in its ability to model the nesting effect of the school environment, thus producing more accurate estimates. The conventional wisdom calls for at least three points of data at level 1. With parallel split scale technique (Lyons et al., 2002), we were able to fit a regression line at level 1 using only pre-test and post-test scores. However, the technique requires enough subscales within each test instrument. This prevents us from using HLM to analyze all of the instruments. In order to be consistent with the results, we chose OLS as the estimating model.

¹¹ We also looked at variables such as student race, limited English proficiency, IEP, attendance rate for teachers and students, and student mobility rate. However, they were excluded from the model because of insufficient or missing data.

2. Is the program equally effective for students from different backgrounds—gender, race, ethnicity, economic status, and English language skills?

The students in our sample are very homogenous in a sense that they are predominantly African Americans and few were characterized as limited English proficiency or requiring an individual education plan. However, we included gender and economic status measures in our models to control for effects from these factors.

3. How does the fidelity of implementation of the Voyager program affect student learning?

We took an approach similar to that for our attribution assessment in studying the effects of level of implementation in the treatment group (see model 2). The main independent variable in this analysis is implementation level rather than Voyager as in model 1. The variable is a sum of four items used in the observer protocol, also referred to as IFI. Each item is rated on a scale of 0 to 3, with 3 indicating a higher level of implementation.¹²

Model 2 (implementation effect)

$$\text{Gain score} = \alpha + \beta \text{ Implement} + \delta_1 \text{Pctmale} + \delta_2 \text{Csize} + \delta_3 \text{Tchexp} + \delta \text{Pctftrl} + \epsilon$$

During the course of the program, each teacher was assessed two or three times by the Voyager program staff. For the purpose of this analysis, we used the final implementation score to describe the level of implementation. These scores were collected prior to data analysis and without knowledge of the evaluation data. In addition, we looked at implementation using both discrete scores and scores grouped into three clusters: high (10–12), medium (7–9), and inadequate (0–6).¹³ The average score for teachers is 9.2. Table 3–3 shows the distribution of implementation scores for teachers.

¹² Different versions of IFI were used in DC and Cleveland. The version used in DC contains four items, while the Cleveland version has seven items, the first four of which are the same as the DC version. This maintains the consistency of scoring between DC and Cleveland by using the four-item scale while taking advantage of the larger sample size. In addition, the evaluator using the seven-item protocol noted that the first four items were most heavily weighted.

¹³ Voyager provides a guarantee of outcomes for students and will not do so where so few features of the program are present.

Table 3–3.—Distribution of implementation scores for teachers

Total scores	Implementation grouping	Number of teachers
12	High	3
11		2
10		
9	Medium	1
8		3
7		
6	Inadequate	1
5		
4		
3		
2		
1		
0		1

NOTE: One teacher did not have implementation scores.

4. What is the relationship between student performance in evaluation test battery and the VIP?

The VIP is the major assessment instrument used for tracking student progress. For the kindergarten assessment, it consists of four subtests: Initial Sound Fluency (ISF), Letter Naming Fluency (LSF), Phoneme Segmentation Fluency (PSF), and Nonsense Word Fluency (NSF). Program managers were interested in examining the relationship between performance on this embedded measure with performance on the assessment battery. Post-test performance on each of these subtests was compared with performance on the tests used in the evaluation study using Pearson correlations.

Findings

This section presents findings for each of the questions.

1. What is the impact of the Voyager Universal literacy System on student literacy learning? Does reading performance in participating schools differ from that in comparable nonparticipating schools in the same district?

1–1. What is the level of achievement for students in Voyager and non-Voyager schools in pre-test and post-test? Tables 3–4 and 3–5 present the descriptive statistics on student achievement

by test instrument. The scores presented in the tables are raw scores, which are the total number of items scored correct for a subset. The value of raw scores is generally limited to research purposes such as making group comparisons or computing correlation coefficients (Wagner, Torgesen, and Rashotte, 1999). For analysis of student gains, we will continue to use raw scores.

Table 3–4.—Descriptive statistics for pre-test student achievement level (raw scores)

Test instrument	Voyager students				Non-Voyager students				Max possible
	Mean	Min	Max	Std dev	Mean	Min	Max	Std dev	
DIBELS Letter Naming Fluency	14.64	0	49	12.30	14.92	0	66	13.64	110
CTOPP Elision	1.11	0	7	1.57	1.06	0	20	2.15	20
CTOPP Blending Words	1.37	0	8	1.96	.89	0	10	1.84	20
CTOPP Blending Nonwords	.59	0	6	.99	.43	0	12	1.24	18
CTOPP Segmenting Words	.22	0	10	1.17	.33	0	18	1.71	20
Woodcock Word Identification	.83	0	51	4.93	1.13	0	58	5.17	64
Woodcock Word Attack	.08	0	7	.56	.28	0	37	2.74	45

Table 3–5.—Descriptive statistics for post-test student achievement level (raw scores)

Test instrument	Voyager students				Non-Voyager students				Max possible
	Mean	Min	Max	Std dev	Mean	Min	Max	Std dev	
DIBELS Letter Naming Fluency	39.39	0	74	14.20	35.05	0	89	18.34	110
CTOPP Elision	3.47	0	13	3.05	2.76	0	18	2.83	20
CTOPP Blending Words	4.89	0	15	3.77	3.14	0	14	3.43	20
CTOPP Blending Nonwords	2.67	0	11	2.47	1.33	0	9	1.97	18
CTOPP Segmenting Words	3.66	0	13	3.96	1.35	0	15	2.98	20
Woodcock Word Identification	9.83	0	52	9.83	8.31	0	61	10.12	64
Woodcock Word Attack	4.73	0	29	5.51	1.34	0	25	3.27	45

In order to understand how the scores for the study students compare to a national sample, we converted the raw scores on the CTOPP to percentiles using the test manuals provided by the publishers (Wagner, Torgesen, and Rashotte, 1999). In addition, we converted the raw scores into standard scores so that raw scores can be compared on a common scale. Percentile and standard scores on a nationally representative sample are not available for the DIBELS. The difficulty of the Woodcock subtests for beginning kindergartners makes percentile conversions quite suspect.¹⁴ The comparisons of means by different types of scores are presented in tables 3–6 and 3–7.

¹⁴ For example, on the Word Identification subtest the mean in the norm sample for October of kindergarten year is “0.” Thus, an average correct score of 1 item can result in highly inflated percentile transformation, suggesting an achievement level for the group which is not representative.

Table 3–6.—Pre-test student CTOPP achievement mean in raw, percentile, and standard scores

Test instrument	Voyager students			Non-Voyager students		
	Raw Score	Percentile ¹	Std Score ²	Raw Score	Percentile	Std Score
CTOPP Elision	1.11	26.32	8.11	1.06	25.72	8.06
CTOPP Blending Words	1.37	29.44	8.37	0.89	24.01	7.89
CTOPP Blending Nonwords	0.59	44.67	9.59	0.43	42.59	9.43
CTOPP Segmenting Words ³	0.27	6.08	5.27	0.33	6.32	5.33

¹ The conversion of pre-test scores for CTOPP used ages 5–0 through 5–5.

² Both percentiles and standard scores were extrapolated. CTOPP standard scores are based on mean of 10 and standard deviation of 3.

³ CTOPP Segmenting Words is normed by students aged 7–0 through 7–5 where the test is normally administered.

Table 3–7.—Post-test student CTOPP achievement mean in raw, percentile, and standard scores

Test instrument	Voyager students			Non-Voyager students		
	Raw Score	Percentile ¹	Std Score ²	Raw Score	Percentile	Std Score
CTOPP Elision	3.47	43.11	9.47	2.76	34.12	8.76
CTOPP Blending Words	4.89	50.00	10.00	3.14	38.82	9.14
CTOPP Blending Nonwords	2.67	58.71	10.67	1.33	41.29	9.33
CTOPP Segmenting Words ³	3.67	16.00	7.00	1.35	9.00	6.00

¹ The conversion of pre-test scores for CTOPP used ages 5–0 through 5–5.

² Both percentiles and standard scores were extrapolated. CTOPP standard scores are based on mean of 10 and standard deviation of 3.

³ CTOPP Segmenting Words is normed by students aged 7–0 through 7–5 where the test is normally administered.

These data show that Voyager and comparison students generally show initial performance that is well below that of the norm group. By the end of the year, with the exception of CTOPP Segmenting Words, which is a very challenging test for kindergarten students, the Voyager students are performing around the national average. The scores of comparison students also improve, but generally remain well below average.

1–2. What is the extent of gains for students in Voyager schools? Westat tested the pre-intervention literacy abilities of students in October 2002 to ascertain the comparability between treatment and comparison groups in initial performance on the evaluation’s test battery. Results of these tests indicate that at the outset of the school year, there were no differences either overall or within the districts between the treatment and comparison groups on the test battery.¹⁵ Overall, the students exhibited very limited prereading skills. On most of the assessments, the vast majority of

¹⁵ Details results were presented in Westat’s January 2003 “Report on the Comparability of Voyager and Non-Voyager Schools.”

students were either screened out¹⁶ or failed to get any items correct when assessed. (The only test on which students were able to have some success was the DIBELS Letter Naming Fluency.)

In conjunction with the post-test results, we found that overall gains for students in Voyager schools were large and significant. The differences between pre-test and post-test were statistically significant at the .00 level on all seven assessments. The effect sizes¹⁷ were very large, ranging from 1.51 to 8.3, which means that students in Voyager schools gained 1.51 to 8.3 standard deviations from September 2002 to May 2003 (table 3–8).

Table 3–8.—Student progress in Voyager schools

Test instrument	Mean	Min	Max	Sd	Sig. (p value)	Effect size
DIBELS Letter Naming Fluency	24.75	-16	65	14.10	.00	2.01
CTOPP Elision	2.35	-3	12	2.61	.00	1.51
CTOPP Blending Words	3.51	-4	13	3.73	.00	1.79
CTOPP Blending Nonwords	2.08	-3	9	2.46	.00	2.10
CTOPP Segmenting Words	3.44	-3	13	3.91	.00	2.94
Woodcock Word Identification	9.00	0	40	8.33	.00	1.83
Woodcock Word Attack	4.65	0	25	5.36	.00	8.30

1–3. Does student reading performance in participating schools differ from that in comparable non-participating schools at district and school level? The second analysis compared the gains between the treatment and comparison group. Gains are presented overall (table 3–9) and by district (tables 3–10 and 3–11).

¹⁶ Depending on the test instruments, a student would be screened out if he/she failed to answer one question correctly on a predetermined number of questions in a row. The score for a screened out student is coded as “0.”

¹⁷ Effect size is the standardized mean difference between treatment and comparison groups or pre-test post-test outcomes. The formula is effect size = $(U_1 - U_2) / \sigma$, where $U_1 - U_2$ is the expected mean difference between groups, σ is the expected variance within groups.

Table 3–9.—Comparison of progress between students in treatment and comparison schools (overall)

Test instrument	Treatment (N=202)	Comparison (N=196)	T-C	Sig. (p value)	Effect Size
DIBELS Letter Naming Fluency	24.75	20.13	4.62	.00	.30
CTOPP Elision	2.35	1.07	1.38	.01	.61
CTOPP Blending Words	3.51	2.26	1.25	.00	.36
CTOPP Blending Nonwords	2.08	.89	1.19	.00	.61
CTOPP Segmenting Words	3.44	1.02	2.42	.00	.93
Woodcock Word Identification	9.00	7.18	1.82	.03	.23
Woodcock Word Attack	4.65	1.07	3.58	.00	1.32

We found that the gains in the treatment schools were larger than those in the comparison schools on all seven assessments. The difference of means test suggests that these gains were significant in all seven cases and the effect sizes were large, ranging from 0.3 to 1.32 with an average of 0.62. Although the number of paired schools is relatively small, the power of the analysis is large (0.78).¹⁸

Table 3–10.—Comparison of progress between students in treatment and comparison schools at district level: District 1

Test instrument	Treatment (N=93)	Comparison (N=79)	T-C	Sig. (p value)	Effect Size
DIBELS Letter Naming Fluency	19.84	16.30	3.54	.10	.28
CTOPP Elision	2.29	1.72	.57	.13	.25
CTOPP Blending Words	3.10	3.06	.04	.95	.01
CTOPP Blending Nonwords	1.49	1.18	1.31	.37	.53
CTOPP Segmenting Words	2.96	1.34	1.62	.00	.49
Woodcock Word Identification	9.91	7.77	2.14	.11	.28
Woodcock Word Attack	4.73	1.16	3.57	.00	1.24

In District 1, the gains in the treatment schools also were larger than the comparison schools on all assessments. Although the gains on only two assessments were statistically significant ($\alpha=0.05$), the effect sizes were mostly moderate to large, ranging from 0.01 to 1.24 with an average of 0.44.

¹⁸ The power analysis (Raudenbush, 1997) is a proxy estimate because the original formula is intended for randomized experiment. The estimate is based on the assumption that the intracluster correlation=0.01, the design $n=57$, $j=4$, and the estimate $d=0.62$.

Table 3–11.—Comparison of progress between students in treatment and comparison schools at district level: District 2

Test instrument	Treatment (N=109)	Comparison (N=117)	T-C	Sig. (p value)	Effect Size
DIBELS Letter Naming Fluency	28.95	22.71	6.24	.00	.39
CTOPP Elision	2.40	1.68	.72	.03	.32
CTOPP Blending Words	3.87	1.72	2.15	.00	.71
CTOPP Blending Nonwords	2.57	.70	1.87	.00	1.26
CTOPP Segmenting Words	3.84	.80	3.04	.00	1.54
Woodcock Word Identification	8.22	6.78	1.44	.16	.17
Woodcock Word Attack	4.59	1.00	3.59	.00	1.37

In District 2, gains in the treatment schools were larger than the comparison schools on six of the seven assessments. The gains on six assessments were statistically significant, and the effect sizes were moderate to large, ranging from 0.17 to 1.54 with an average of 0.82.

We further analyzed the difference between treatment and comparison groups at the school level.¹⁹ Tables 3–12 through 3–15 show that in three out of four pairs, the Voyager students clearly outperformed their counterparts. In the fourth pair, the comparison school (1–1C) showed higher gains on five of the seven assessments.

Table 3–12.—Comparison of progress between students in treatment and comparison schools at school level: District 1, school pair 1

Test instrument	1–1T (N=55)	1–1C (N=30)	Difference
DIBELS Letter Naming Fluency	19.05	15.00	4.05
CTOPP Elision	1.89	1.91	-.02
CTOPP Blending Words	1.91	3.37	-1.46
CTOPP Blending Nonwords	.91	1.43	-.52
CTOPP Segmenting Words	1.75	2.03	-.28
Woodcock Word Identification	7.24	8.17	-.93
Woodcock Word Attack	2.73	1.47	1.26

¹⁹ For this set of comparisons, we removed the district and school identifiers.

Table 3–13.—Comparison of progress between students in treatment and comparison schools at school level: District 1, school pair 2

Test instrument	1–2T (N=38)	1–2C (N=49)	Difference
DIBELS Letter Naming Fluency	20.97	17.10	3.87
CTOPP Elision	2.87	1.61	1.26
CTOPP Blending Words	4.82	2.88	1.94
CTOPP Blending Nonwords	2.34	1.02	1.32
CTOPP Segmenting Words	4.71	.92	3.79
Woodcock Word Identification	7.53	9.84	-2.31
Woodcock Word Attack	7.63	.98	6.65

Table 3–14.—Comparison of progress between students in treatment and comparison schools at school level: District 2, school pair 1

Test instrument	2–1T (N=44)	2–1C (N=69)	Difference
DIBELS Letter Naming Fluency	26.48	21.04	5.44
CTOPP Elision	2.18	1.87	.31
CTOPP Blending Words	3.95	2.16	1.79
CTOPP Blending Nonwords	3.39	.84	2.55
CTOPP Segmenting Words	3.93	.81	3.12
Woodcock Word Identification	9.84	6.54	3.30
Woodcock Word Attack	5.57	1.16	4.41

Table 3–15.—Comparison of progress between students in treatment and comparison schools at school level: District 2, school pair 2

Test instrument	2–2T (N=65)	2–2C (N=48)	Difference
DIBELS Letter Naming Fluency	30.62	25.10	5.52
CTOPP Elision	2.55	1.42	1.13
CTOPP Blending Words	3.82	1.08	2.64
CTOPP Blending Nonwords	2.03	.50	1.53
CTOPP Segmenting Words	3.78	.79	2.99
Woodcock Word Identification	7.12	7.13	-.01
Woodcock Word Attack	3.92	.77	3.15

1–4. To what extent can the gains be attributed to the program? We conducted multiple regression analyses to assess the extent of which the observed gains are attributed to the Voyager program. Table 3–16 presents a summary of the regression results for the program effect models, while more detailed results are contained in Appendix A. Our analyses suggest that controlling for other variables, the Voyager program has a statistically significant effect ($p \leq .05$) on student learning on six of the seven assessments. In the case of DIBELS, for example, students in Voyager schools gained 5 points more on average than their counterparts in the comparison schools. Other variables are significant on isolated assessments, and there is no consistent pattern across instruments. Further, we found that the Voyager program has a stronger effect than any student and teacher characteristics examined in the sample (see standardized coefficients in Appendix A).

Table 3–16.—Regression summary of the program effect models (N=398)

Variable	DIBELS	CTOPP Elision	CTOPP Blending Words	CTOPP Blending Nonwords	CTOPP Segmenting Words	Woodcock Word Identification	Woodcock Word Attack
Voyager	5.00***	.69**	---	1.40***	2.29***	2.30**	3.94***
Percent male	-.34**	---	---	---	---	---	---
Class size	---	---	---	-.05*	---	---	---
Teacher experience	---	---	.07**	.05***	---	---	.11***
Percent free lunch	---	---	---	.02**	---	---	---

*($p \leq .10$), **($p \leq .05$), ***($p \leq .01$).

NOTE: Only statistically significant results were reported. Statistics presented are unstandardized coefficients.

2. Is the program equally effective for students from different backgrounds—gender, race, ethnicity, economic status, and English language skills?

The homogenous nature of the sample does not allow us to investigate the differential impacts of the program for students from different background. However, we did control for gender and economic status in our model. Table 3–16 suggests little consistent pattern of effects across instruments.

3. How does the fidelity of implementation of the Voyager program affect student learning?

We found that higher level of teacher implementation has a positively significant effect ($p \leq .05$) on student achievement on all seven assessments (table 3–17; see Appendix B for more detailed results). Further results show that, for example, one additional point in teacher implementation score contributes to a 1.8 point gain in students’ DIBELS scores. Teacher experience, class size,

and percent of students receiving free lunch are significant in more than three instruments, which mean that in Voyager classes, smaller class size, lower percentage of students receiving free or reduced-price lunch, and more experienced teachers are associated with larger gains in scores. However, the Voyager program has a stronger effect than any other student and teacher characteristics in the sample (see standardized coefficients in Appendix B).

**Table 3–17.—Regression summary of the program implementation models
(raw scores, N=202)**

Variable	DIBELS	CTOPP Elision	CTOPP Blending Words	CTOPP Blending Nonwords	CTOPP Segmenting Words	Woodcock Word Identification	Woodcock Word Attack
Implement	1.80**	.41***	.96***	.63***	.78***	2.21***	1.45***
Percent male	-.87***	---	---	---	---	-.41**	---
Class size	---	---	.60**	-.59***	-.68**	---	-1.07**
Teacher experience	---	---	---	.08**	.12*	---	.19*
Percent free lunch	---	---	-.12***	-.07**	-.13***	-.28**	-.22***

*($p \leq .10$), **($p \leq .05$), ***($p \leq .01$).

NOTE: Only statistically significant results were reported. Statistics presented are unstandardized coefficients.

Finally, table 3–18 presents the results of implementation models based on the grouping of teachers’ implementation scores. The grouping proved to be statistically significant on all seven assessments. For example, in DIBELS, students from high implementation classrooms gained an average of 6.99 raw score points more than those from medium implementation classrooms and 13.98 raw score points more than their counterparts in low implementation classrooms. Table 3–19 shows pre- and post-changes by implementation level for the remaining assessments using raw scores. While the limited sample size suggests these data should be used with caution, the data show greater gains for high implementation classrooms.

Table 3–18.—Regression summary of the program implementation models (grouping, N=202)

Variable	DIBELS	CTOPP Elision	CTOPP Blending Words	CTOPP Blending Nonwords	CTOPP Segmenting Words	Woodcock Word Identification	Woodcock Word Attack
Implement	6.99**	1.64***	3.63***	2.41***	2.86***	8.44***	5.56***
Percent male	-.70**	---	---	---	---	---	---
Class size	---	---	-.59**	-.59***	-.66**	---	-1.07**
Teacher experience	---	---	---	.09**	.14**	---	.24**
Percent free lunch	---	---	-.14***	-.08**	-.14**	-.32**	-.24***

*(p ≤ .10), ***(p ≤ .05), ***(p ≤ .01).

NOTE: Only statistically significant results were reported. Statistics presented are unstandardized coefficients.

Table 3–19.—Pre and post raw scores by implementation level

Test instrument	High ¹			Medium ²			Inadequate ³		
	Pre	Post	Δ	Pre	Post	Δ	Pre	Post	Δ
DIBELS Letter Naming Fluency	15.71	42.61	26.90	11.54	39.68	28.14	13.88	36.13	22.25
CTOPP Elision	1.38	4.18	2.80	1.00	2.90	1.90	1.08	3.65	2.57
CTOPP Blending Words	1.60	6.29	4.69	.66	3.44	2.78	1.13	4.15	3.02
CTOPP Blending Nonwords	.61	3.61	3.00	.36	1.76	1.40	.45	2.15	1.70
CTOPP Segmenting Words	.19	4.29	4.10	.30	2.50	2.20	.18	3.68	3.50
Woodcock Word Identification	.81	12.33	11.52	.16	7.44	7.28	1.78	9.00	7.22
Woodcock Word Attack	.14	6.96	6.82	.00	2.54	2.54	.13	3.50	3.37

¹ Five teachers and 79 students.

² Three teachers and 50 students.

³ Two teachers and 40 students.

4. What is the relationship between student performance in evaluation test battery and the VIP?

Examination of student performance on the evaluation assessment battery and the VIP shows that the VIP has moderate and significant correlations with almost all of the evaluation test battery (table 3–20). One might question, however, whether or not a .59 is as high as might be expected for the DIBELS Letter Naming Fluency assessment that is included in both the VIP and the evaluation test battery.

Table 3–20.—Correlations between evaluation test battery and the VIP

VIP	DIBELS Letter Naming Fluency	CTOPP Elision	CTOPP Blending Words	CTOPP Blending Nonwords	CTOPP Segmenting Words	Woodcock Word Identification	Woodcock Word Attack
Initial Sound Fluency (N=102)	.23**	.23**	.27**	.17	.40**	.46**	.40**
Letter Naming Fluency (N=198)	.59**	.39*	.50**	.35**	.36**	.53**	.53**
Phoneme Segmentation Fluency (N=198)	.29**	.43**	.51**	.37**	.43**	.39**	.44**
Nonsense Word Fluency (N=198)	.42**	.39**	.48**	.31**	.33**	.60**	.56**

*($p \leq .05$), **($p \leq .01$).

4. CONCLUSIONS

This 1-year study of kindergarten students in Voyager and comparable non-Voyager classrooms provides strong evidence for the effectiveness of the Voyager program in providing students in urban environments with early reading skills. Overall, the students using the Voyager program showed significantly greater gains in skills such as letter naming fluency, elision, blending words and nonwords, segmenting words, word attack, and word identification. The analyses also show that the differences between groups were large enough to represent meaningful differences, with effect sizes ranging from 0.23 to 1.32 with a mean of 0.62.²⁰

Implementation data collected by Voyager program staff provide additional information of importance. These data show that when the program is implemented with greater fidelity, greater gains are found. Students from high implementation classrooms scored on average 6.99 points higher on the DIBELS than those from classrooms judged to be medium implementers and 13.98 points higher than those in classrooms with low implementation scores. The Voyager students failed to outperform the comparison students in only one pair of schools. There, the Voyager school had been judged as having very incomplete implementation and was considered a problematic example of the Voyager approach.

The implementation findings also show some interactions worth noting. While the Voyager program has a stronger effect than any other student and teacher characteristics in the sample, on four out of seven tests, there is a significant negative effect of class size, meaning that students in smaller classes outperformed those in larger classes; on five out of seven tests, classes with smaller numbers of students receiving free or reduced-price lunch outperform those with larger number of such students; and finally, on three out of seven tests, students of teachers with greater experience outperformed those of teachers with less experience. While these findings are far from unusual, they do present challenges to the program's goal of serving all students well. A closer look at how the program may be implemented differently in these situations is recommended to see whether there are additional supports or strategies that the Voyager program might provide.

The data provide strong testimony that Voyager can be highly effective with African American students in urban environments. While the study was small, the results are strong. The Voyager program appears to provide a very solid foundation for successful literacy development.

²⁰ In assessing these findings, it is important to keep in mind that initial tests of comprehension were not included in this study. A critical question is whether the advantages found here persist as comprehension increases in importance across the grade levels.

REFERENCES

- Good, R.H., and Kaminski, R.A. (Eds.). (2002). *Dynamic Indicators of Basic Early Literacy Skills*, 6th ed. Eugene, OR: Institute for Development of Educational Achievement.
- Good, R.H., Wallin, J.U., Simmons, D.C., Kame'enui, E.J., and Kaminski, R.A. (2002). *System-wide Percentile Ranks for DIBELS Benchmark Assessment* (Technical Report No. 9). Eugene, OR: University of Oregon.
- Lyons, K.S., Zarit, S.H., Sayer, A.G., and Whitlach, C.J. (2002). Caregiving as a Dyadic Process: Perspectives from the Caregiver and Receiver. *Journal of Gerontology*, 57(3), 195-204.
- Raudenbush, S. (1997). Statistical Analysis and Optimal Design for Cluster Randomized Trials. *Psychological Methods*, 2(2):173-185.
- Wagner, R., Torgesen, J., and Bryant, D. (1999). *Comprehensive Test of Phonological Processing*. Austin, TX: PRO-ED, Inc.
- Wagner, R.K., Torgesen, J.K., and Rashotte, C.A. (1999). *CTOPP Comprehensive Test of Phonological Processing Examiner's Manual*. Austin, TX: PRO-ED, Inc.
- Woodcock, R. (1998). *Woodcock Reading Mastery Tests-Revised*. Circle Pines, MN: American Guidance Service.
- Woodcock, R.W. (1998). *Woodcock Reading Mastery Tests-Revised (Forms G and H) Examiner's Manual*. Circle Pines, MN: American Guidance Service.

Appendix A.
Detailed Regression Results for Program Effects

Table A-1.—Regression results for program effect (DIBELS Letter Naming Fluency)

Variable	Unstd. coeff.	Std. error	Std. coeff.	T	Sig
Constant	26.03	10.87	---	2.40	.02
Voyager	5.00	1.91	.17	2.62	.01
Percent Male	-.34	.13	-.16	-2.46	.01
Class size	.11	.21	.03	.52	.60
Teacher experience	.04	.13	.02	.30	.76
Percent free lunch	.07	.06	.09	1.17	.25

Adj R² = .04, F (sig) = .01

Table A-2.—Regression results for program effect (CTOPP Elision)

Variable	Unstd. coeff.	Std. error	Std. coeff.	T	Sig
Constant	.28	1.82	---	.15	.88
Voyager	.69	.32	.14	2.15	.03
Percent Male	.00	.02	.00	-.02	.99
Class size	.05	.04	.09	1.41	.16
Teacher experience	.00	.02	-.01	-.11	.91
Percent free lunch	.01	.01	.04	.56	.58

Adj R² = .01, F (sig) = .27

Table A-3.—Regression results for program effect (CTOPP Blending Words)

Variable	Unstd. coeff.	Std. error	Std. coeff.	T	Sig
Constant	6.12	2.68	---	2.29	.02
Voyager	.65	.47	.09	1.38	.17
Percent Male	-.05	.03	-.10	-1.58	.12
Class size	-.03	.05	-.03	-.56	.57
Teacher experience	.07	.03	.13	2.20	.03
Percent free lunch	-.01	.01	-.06	-.81	.42

Adj R² = .02, F (sig) = .04

Table A-4.—Regression results for program effect (CTOPP Blending Nonwords)

Variable	Unstd. coeff.	Std. error	Std. coeff.	T	Sig
Constant	.53	1.67	---	.32	.75
Voyager	1.40	.29	.30	4.77	.00
Percent Male	-.01	.02	-.04	-.67	.50
Class size	-.05	.03	-.10	-1.65	.10
Teacher experience	.05	.02	.14	2.48	.01
Percent free lunch	.02	.01	.16	2.29	.02

Adj R²=.1, F (sig)=.00**Table A-5.—Regression results for program effect (CTOPP Segmenting Words)**

Variable	Unstd. coeff.	Std. error	Std. coeff.	T	Sig
Constant	3.02	2.46	---	1.23	.22
Voyager	2.29	.43	.33	5.30	.00
Percent Male	-.03	.03	-.07	-1.05	.30
Class size	-.04	.05	-.04	-.74	.46
Teacher experience	.03	.03	.05	.88	.38
Percent free lunch	.00	.01	.00	.03	.98

Adj R²=.1, F (sig)=.00**Table A-6.—Regression results for program effect (Woodcock Word Identification)**

Variable	Unstd. coeff.	Std. error	Std. coeff.	T	Sig
Constant	14.71	6.37	---	2.31	.02
Voyager	2.30	1.12	.13	2.05	.04
Percent Male	-.12	.08	-.10	-1.59	.11
Class size	.09	.12	.04	.69	.49
Teacher experience	.11	.08	.08	1.42	.16
Percent free lunch	-.05	.03	-.11	-1.46	.15

Adj R²=.04, F (sig)=.01**Table A-7.—Regression results for program effect (Woodcock Word Attack)**

Variable	Unstd. coeff.	Std. error	Std. coeff.	T	Sig
Constant	2.00	3.26	---	.61	.54
Voyager	3.94	.57	.41	6.87	.00
Percent Male	-.06	.04	-.08	-1.40	.16
Class size	.05	.06	.05	.84	.40
Teacher experience	.11	.04	.15	2.74	.01
Percent free lunch	.00	.02	-.01	-.08	.94

Adj R²=.19, F (sig)=.00

Appendix B.
Detailed Regression Effects for Implementation Effects

Table B–1.—Regression results for the implementation (DIBELS Letter Naming Fluency)

Variable	Unstd. coeff.	Std. error	Std. coeff.	T	Sig
Constant	73.41	24.19	---	3.03	.00
Implement	1.80	.83	.33	2.16	.03
Percent Male	-.87	.31	-.34	-2.73	.01
Class size	-.82	1.02	-.20	-.80	.43
Teacher experience	-.20	.24	-.11	-.84	.40
Percent free lunch	-.12	.19	-.16	-.62	.54

Adj R² = .11, F (sig) = .00

Table B–2.—Regression results for the implementation (CTOPP Elision)

Variable	Unstd. coeff.	Std. error	Std. coeff.	T	Sig
Constant	4.01	4.51	---	.89	.38
Implement	.41	.16	.42	2.67	.01
Percent Male	-.05	.06	-.10	-.79	.43
Class size	-.05	.19	-.07	-.29	.77
Teacher experience	-.01	.04	-.04	-.28	.78
Percent free lunch	-.03	.04	-.23	-.84	.40

Adj R² = .06, F (sig) = .03

Table B–3.—Regression results for the implementation (CTOPP Blending Words)

Variable	Unstd. coeff.	Std. error	Std. coeff.	T	Sig
Constant	21.46	6.26	---	3.43	.00
Implement	.96	.22	.66	4.44	.00
Percent Male	-.13	.08	-.19	-1.56	.12
Class size	-.60	.26	-.53	-2.28	.03
Teacher experience	.06	.06	.12	.98	.33
Percent free lunch	-.12	.05	-.67	-2.65	.01

Adj R² = .18, F (sig) = .00

Table B-4.—Regression results for the implementation (CTOPP Blending Nonwords)

Variable	Unstd. coeff.	Std. error	Std. coeff.	T	Sig
Constant	14.67	3.87	---	3.79	.00
Implement	.63	.13	.64	4.72	.00
Percent Male	-.04	.05	-.09	-.82	.41
Class size	-.59	.16	-.77	-3.62	.00
Teacher experience	.08	.04	.22	1.97	.05
Percent free lunch	-.07	.03	-.55	-2.39	.02

Adj R² = .32, F (sig) = .00**Table B-5.—Regression results for the implementation (CTOPP Segmenting Words)**

Variable	Unstd. coeff.	Std. error	Std. coeff.	T	Sig
Constant	17.67	6.59	---	2.68	.01
Implement	.78	.23	.53	3.42	.00
Percent Male	.01	.09	.02	.12	.90
Class size	-.68	.28	-.60	-2.45	.02
Teacher experience	.12	.07	.23	1.80	.08
Percent free lunch	-.13	.05	-.70	-2.61	.01

Adj R² = .09, F (sig) = .01**Table B-6.—Regression results for the implementation (Woodcock Word Identification)**

Variable	Unstd. coeff.	Std. error	Std. coeff.	T	Sig
Constant	46.27	15.61	---	2.96	.00
Implement	2.21	.54	.62	4.12	.00
Percent Male	-.41	.20	-.25	-2.05	.04
Class size	-.82	.66	-.29	-1.24	.22
Teacher experience	-.02	.15	-.02	-.14	.89
Percent free lunch	-.28	.12	-.60	-2.32	.02

Adj R² = .16, F (sig) = .00**Table B-7.—Regression results for the implementation (Woodcock Word Attack)**

Variable	Unstd. coeff.	Std. error	Std. coeff.	T	Sig
Constant	27.83	9.95	---	2.80	.01
Implement	1.45	.34	.64	4.25	.00
Percent Male	-.01	.13	-.01	-.11	.91
Class size	-1.07	.42	-.61	-2.55	.01
Teacher experience	.19	.10	.24	1.94	.06
Percent free lunch	-.22	.08	-.73	-2.85	.01

Adj R² = .15, F (sig) = .00