

Evaluation of the Voyager Universal Literacy System® Year 2

August 2004

Authors:

Dr. Joy Frechtling
Dr. Xiaodong Zhang
Dr. Lawrence Wang

Prepared for:

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

By



TABLE OF CONTENTS

Chapter	Page
ABSTRACT	4
IMPACT OF VOYAGER ON STUDENT ACHIEVEMENT	5
Evaluation Questions	5
Study Activity	6
Attrition of Sample	6
Assessment Battery	6
Implementation Measures	8
Analysis and Findings	8
Methodological Exploration	22
CONCLUSION	23
REFERENCES	24

List of Appendices

Appendix	Page
A Detailed HLM results for program effects	25
B Detailed HLM results for implementation effects	31
C Using HLM to Analyze Pre-Test and Post-Test Scores	37

TABLE OF CONTENTS—continued

List of Tables

Table		Page
1	Change of Study Sample from T2 to T3	6
2	Full Test Battery	7
3	Number of Students for Assessment by Wave	8
4	Descriptive Statistics of Student Achievement (T3: raw score)	9
5	Student Achievement in Comparison to National Average (T3: CTOPP)	10
6	Comparison of progress (T3 - T1) between Students in Voyager and Non-Voyager Schools: Overall	10
7	Comparison of progress (T3 - T1) between students in Voyager and Non-Voyager schools: District level	11
8	Comparison of progress (T3 - T1) between students in Voyager and Non-Voyager schools: School level	12
9	Comparison of Raw Scores between Voyager and Non-Voyager Students (T1 to T3)	13
10	HLM results for the program effect	17
11	OLS results for the program effect	18
12	Descriptive Statistics for Classroom Characteristics (T3)	19
13	Distribution of Implementation Scores for Teachers (T3)	20
14	HLM results for the implementation effect	21
15	OLS results for the implementation effect	21
16	Comparison of Student Raw Scores between High and Low Implementation Classes (T3)	22

TABLE OF CONTENTS—continued

List of Figures

Figure		Page
1	Comparison of growth in CTOPP-Elision	13
2	Comparison of growth in CTOPP-Blending Words	13
3	Comparison of growth in CTOPP-Blending Nonwords	14
4	Comparison of growth in CTOPP-Segmenting Words	14
5	Comparison of growth in Woodcock Word Identification	15
6	Comparison of growth in Woodcock Word Attack	15

List of Models

Model		Page
1	Program Effect (HLM)	16
2	Program Effect (OLS)	18
3	Implementation Effect (HLM)	20
4	Implementation Effect (OLS)	21

ABSTRACT

In 2002, Westat began a study of early literacy attainment in two inner city school districts, comparing the reading achievement of kindergarten students participating in the Voyager Universal Literacy System with that of non-participants in a matched sample of schools. The study found that kindergarten students in the Voyager program showed greater learning gains than the comparison students on seven individually-administered assessments of early reading skills. Further, gains were related to program implementation; schools showing greater fidelity in implementation also showed larger gains.

This report presents the results of a second year of the evaluation of the Voyager Universal Literacy System, following 255 kindergarten students in 8 schools through their first grade year. The second year of data collection provides an opportunity to conduct a more rigorous longitudinal examination of the program impacts on student achievement. The findings again show that students participating in the Voyager program increased their reading skills more than students in the comparison schools. Performance significantly favored the Voyager students on eight out of nine assessments administered. On the measure with no significant difference, most students in both treatment and comparison schools correctly answered all of the questions in the test. In addition, the difference in achievement progress between the two groups is large, with effect sizes ranging from 0.42 to 1.08 across assessment instruments. Further, the average scores of the Voyager students at the end of first grade were largely at or above the national average, while those of comparison students remained below the national average.

Using hierarchical linear modeling and ordinary least square methods, we found that the Voyager program has statistically significant positive impacts on student achievement in seven out of nine assessments. Looking at classroom characteristics, the percentage of males and students with free and reduced price lunch attenuated the effects; classrooms with larger percentages of males and/or students receiving free and reduced price lunch made smaller gains. Fidelity of implementation, as measured by the Instructional Fidelity Index, continued to be significantly related to learning on five of the nine measures.

IMPACT OF VOYAGER ON STUDENT ACHIEVEMENT

In 2002, Voyager Expanded Learning and the Council of Great City Schools contracted with Westat to conduct an evaluation of its K–3 elementary program. The first year evaluation provided evidence of the success of the Voyager Universal Literacy System on student literacy for primarily African American kindergarten students in urban schools (Frechtling, Silverstein, and Zhang, 2003). This report presents findings from the second year of the evaluation, following students in the Voyager and comparison schools into their first grade year.

This second year of data collection provides an opportunity to conduct a broader and more rigorous examination of the program impacts on student achievement. Using six assessments for 2 years, we are able to measure student achievement at three points in time, which allows for a longitudinal analysis by using hierarchical linear modeling (HLM). In addition, comparisons are expanded to include impacts on increasingly advanced skills in early literacy, covering skills expected to be learned at both kindergarten and first grades.

A description of the Voyager program, our approach to evaluation, and the sampling process is presented in our evaluation of the Voyager Universal Literacy System (Frechtling, Silverstein, and Zhang, 2003). In this report, we revisit the evaluation questions guiding the study, discuss study activities in year 2, and present the analysis and findings.

Evaluation Questions

The year 2 evaluation continues to address 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 such as gender, race, ethnicity, economic status, and English language skills?
- How does the level of implementation affect outcomes?

Study Activity

In year 2, grade 1 teachers in the treatment group received training and materials from Voyager in August 2003, and they began implementing the program in early September 2003. In May 2004, Westat administered the common core reading battery to students who had been tested as kindergarteners in October 2002 and May 2003.

Attrition of Sample

Of the 398 students in the year 1 sample, 255 remained at the end of year 2—137 in the treatment group and 118 in the comparison group (table 1). Examination of attrition rates for the two groups shows that the overall attrition rates, though high, are typical for the eight schools, with the rate for the treatment group being slightly lower than that for the comparison group. The difference, however, is not statistically significant. In addition, comparison of student achievement at the end of year 1 (T2) between the stayers and movers show that stayers had statistically significant higher average scores than movers in 3 assessments including CTOPP Elision, CTOPP-Blending Words, and Woodcock Word Attack. But the differences were not significant in DIBELS-Letter Naming Fluency, CTOPP-Blending Nonwords, CTOPP-Segmenting Words, and Woodcock Word Identification.

Table 1.—Change of study sample from T2 to T3

Sample	May 2003 (T2)	May 2004 (T3)	Attrition rate (%)
Total	398	255	35.9
Treatment	202	136	32.2
Comparison	196	118	39.8

Assessment Battery

In selecting the battery to test student literacy in year 2, we tried to maintain consistency with year 1 measures while adding three assessments for a more advanced stage of student reading ability. As a result, of nine assessments administered in year 2, six were the same as in the year 1. Table 2 presents the instruments used in both years with the shaded area indicating that the same instrument was used.

Table 2.—Full test battery

2003 (T1, T2)	2004 (T3)
DIBELS-Letter Naming Fluency	DIBELS-Oral Reading Fluency
CTOPP-Elision	CTOPP-Elision
CTOPP-Blending Words	CTOPP-Blending Words
CTOPP-Blending Nonwords	CTOPP-Blending Nonwords
CTOPP-Segmenting Words	CTOPP-Segmenting Words
Woodcock Word Identification	Woodcock Word Identification
Woodcock Word Attack	Woodcock Word Attack
	WRAT3-Name/Letter Writing
	WRAT3-Spelling

In selecting the battery to test student literacy in year 2, we tried to maintain consistency with year 1 measures while adding three assessments for a more advanced stage of student reading ability. As a result, of nine assessments administered in year 2, six were the same as in the year 1. Table 2 presents the instruments used in both years with the shaded area indicating that the same instrument was used.

- **Dynamic Indicators of Basic Early Literacy Skills (DIBELS)-Oral Reading Fluency subtest.** DIBELS is a standardized, individually administered test. Intended for most children from mid-first grade through third grade, Oral Reading Fluency is a set of passages and procedures designed to (a) identify children who may need additional instructional support, and (b) monitor progress toward instructional goals. Student performance is measured by having students read a passage aloud for 1 minute. Words omitted, words substituted, and hesitations of more than 3 seconds are scored as errors. Words self-corrected within 3 seconds are scored as accurate. The number of correct words per minute from the passage is the oral reading fluency rate.
- **Wide Range Achievement Test—Name/Letter Writing subtest.** The purpose of this subtest is to measure the codes that are needed to learn the basic skills of reading. Administered individually, the subtest contains 15 letters of alphabet.
- **Wide Range Achievement Test—Spelling subtest.** The purpose of this subtest is to measure the codes that are needed to learn the basic skills of spelling. Administered individually, the subtest assesses the pronunciation of 42 words. After 10 consecutive errors on the test, the examiner should terminate the administration of the formal items of that subtest.

This study presents findings for only students tested in three assessment periods in fall 2002, spring 2003, and spring 2004. Table 3 shows the number of students with valid scores for each assessment by wave and program status¹.

Table 3.—Number of students for assessment by wave

Test instrument	T1		T2		T3	
	Voyager	Non-Voyager	Voyager	Non-Voyager	Voyager	Non-Voyager
DIBELS-Letter Naming Fluency	229	215	202	196	NA	NA
DIBELS-Oral Reading Fluency	NA	NA	NA	NA	136	118
CTOPP-Elision	231	216	202	196	137	118
CTOPP-Blending Words	227	216	201	192	136	117
CTOPP-Blending Nonwords	225	216	202	196	137	118
CTOPP-Segmenting Words	222	216	202	196	137	118
Woodcock Word Identification	231	214	202	196	137	118
Woodcock Word Attack	230	215	202	196	137	118
WRAT3-Name/Letter Writing	NA	NA	NA	NA	137	118
WRAT3-Spelling	NA	NA	NA	NA	137	118

Implementation Measures

In addition to data on student achievement, the study drew on data on the program implementation from the Instructional Fidelity Index (IFI) gathered by Voyager training staff. A detailed description of IFI is presented in year 1 report.

Analysis and Findings

In this section, we discuss the analytical approaches and findings for each evaluation question. We have organized findings so that they parallel those presented in the year 1 report.

- 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 non-participating schools in the same district?**

¹ Variation in numbers was caused by refusals to participate, or error in test administration.

We break down this question into three subquestions, which we present and discuss in the pages that follow.

1-1. What is the level of achievement for students in Voyager and non-Voyager schools in May 2004?

Table 4 presents the descriptive statistics on student achievement by test instrument at the end of grade 1. The scores presented in the tables are raw scores, which are the total number of items scored correctly for the assessment. It is important to note that number of questions for each assessment varies from 15 to 200,² and therefore, the achievement level and gain should be viewed in the context of each assessment and not compared across assessments. The data show that on average, Voyager students have statistically significant higher scores than the non-Voyager students in eight out of nine assessments. Only on WRAT3-Name/Letter Writing did students in Voyager schools fail to show a significant difference compared to their counterparts in the comparison schools. Because the subtest only contains 15 items and students in both groups had average scores of over 14.6, it appears that this test was too easy and a “ceiling effect” has taken place.

Table 4.—Descriptive statistics of student achievement (T3: raw score)

Test instrument	Voyager students (N=137)		Non-Voyager students (N=118)		Voyager-non-Voyager difference
	Mean	Std dev	Mean	Std dev	Sig
DIBELS-Oral Reading Fluency	40.61	26.28	33.05	24.14	0.02
CTOPP-Elision	6.92	3.54	5.36	3.25	0.00
CTOPP-Blending Words	10.34	3.37	6.42	2.82	0.00
CTOPP-Blending Nonwords	5.56	3.17	2.57	2.17	0.00
CTOPP-Segmenting Words	8.18	4.21	3.82	3.86	0.00
Woodcock Word Identification	37.15	15.03	31.54	13.84	0.00
Woodcock Word Attack	16.29	10.35	6.85	7.62	0.00
WRAT3-Name/Letter Writing	14.61	0.98	14.63	0.87	0.91
WRAT3-Spelling	5.21	2.98	4.18	2.67	0.00

In order to understand how a particular score compares in a national sample, we converted individual student raw scores at T3 to percentile scores, using CTOPP as an example. Looking at the distribution of percentile scores, we found that the majority of Voyager students score at or above national average, while the majority of non-Voyager students score below the national average at the end of grade 1 (table 5).

² CTOPP Elision (20), CTOPP Blending Words (20), CTOPP Blending Nonwords (18), CTOPP Segmenting Words (20), Woodcock Word Identification (64), Woodcock Word Attack (45), DIBELS-Oral Reading Fluency (200), WRATS-Name/Letter Writing (15), WRAT3-Spelling (40).

Table 5.—Student achievement in raw and percentile scores using an average score approach

Test instrument	Voyager students (N=137)		Non-Voyager students (N=118)	
	Percent at or above national average	Percent below national average	Percent at or above national average	Percent below national average
CTOPP-Elision	70.1	29.9	50.8	49.2
CTOPP-Blending Words	78.8	20.4	31.4	67.8
CTOPP-Blending Nonsense Words	58.4	41.6	20.3	79.7
CTOPP-Segmenting Words	48.9	51.1	17.0	83.0

Note: The conversion of scores used age groups 6-6 through 6-11. CTOPP segmenting words is normed by students aged 7-0 through 7-5, to whom the test is usually administered. Percents may not add to 100 due to missing values.

1-2. Does student reading performance in Voyager schools differ from those in non-Voyager schools?

In addition to comparing student achievement status, we compared the average achievement gains from October 2002 to May 2004 between treatment and comparison schools using independent sample t-tests. Only student achievement from the six instruments used at both T1 and T3 were analyzed. We found that the gains in the treatment schools are larger than those in the comparison schools on all six assessments. The difference of means test suggests that these gains are statistically significant. The effect sizes across assessment instruments³ range from medium (0.42) to large (1.08) with an average of 0.82⁴ (table 6).

Table 6.—Comparison of progress (T3 to T1) between students in Voyager and non-Voyager schools: Overall

Test instrument	Voyager	Non-Voyager	Voyager-non-Voyager difference	Significance	Effect size
CTOPP-Elision	5.80	4.45	1.35	.00	0.45
CTOPP-Blending Words	8.89	5.57	3.32	.00	0.94
CTOPP-Blending Nonwords	4.93	2.21	2.72	.00	1.01
CTOPP-Segmenting Words	7.99	3.64	4.35	.00	1.08
Woodcock Word Identification	36.26	30.50	5.76	.00	0.42
Woodcock Word Attack	16.20	6.72	9.48	.00	1.07

³ The effect size estimates are standardized mean differences (d), based on comparison between treatment and comparison groups (Lipsey and Wilson, 2001).

⁴ Normally, effect size estimate is instrument-specific and should not be averaged. It is done here only for illustrative purpose.

Table 7 shows district-level comparisons. In district 1, the gains in the treatment schools are larger than the comparison schools on all six assessments. The differences are statistically significant in all assessments and the average effect size is 0.65. The results for district 1 are stronger than in year 1, where only two assessments were statistically significant and the average effect size was 0.44.

In district 2, the gains in the treatment schools are also larger than the comparison schools on all six assessments. The differences are statistically significant in all assessments, and the average effect size is 1.03, which is also greater than year 1.

Table 7.—Comparison of progress (T3 to T1) between students in Voyager and non-Voyager schools: District level

Test instrument	Voyager	Non-Voyager	Voyager-non-Voyager difference	Significance	Effect size
District 1	(N=66)	(N=53)			
CTOPP-Elision	6.26	4.94	1.32	0.02	0.43
CTOPP-Blending Words	8.35	6.06	2.29	0.00	0.63
CTOPP-Blending Nonwords	4.75	2.23	2.52	0.00	0.94
CTOPP-Segmenting Words	8.17	4.57	3.60	0.00	0.81
Woodcock Word Identification	38.89	34.22	4.67	0.08	0.33
Woodcock Word Attack	17.73	10.004	7.73	0.00	0.78
District 2	(N=71)	(N=65)			
CTOPP-Elision	5.37	4.05	1.32	0.01	0.45
CTOPP-Blending Words	9.42	5.17	4.25	0.00	1.25
CTOPP-Blending Nonwords	5.10	2.20	2.90	0.00	1.07
CTOPP-Segmenting Words	7.82	2.89	4.93	0.00	1.40
Woodcock Word Identification	33.82	27.46	6.36	0.01	0.47
Woodcock Word Attack	14.79	4.05	10.74	0.00	1.51

Table 8 presents the differences in student progress between treatment and comparison groups at the school level, and Voyager students clearly have made greater progress than their counterparts in all four school pairs. Only on one assessment in one school pair did the students in comparison schools have larger gains. In addition, Voyager students appear to attain higher levels of gains in year 2 than the previous year. We also calculated the p value and effect size. It is important to note that as the sample size for each pair is small, it is difficult to attain the significance level. There are differences among school pairs in program effect size, with the average effect size being 0.45 (pair 1), 1.17 (pair 2), 1.20 (pair 3), and 1.07 (pair 4).

**Table 8.—Comparison of progress (T3 to T1) between students in Voyager and non-Voyager schools:
School level**

Test instrument	Voyager	Non-Voyager	Voyager-non-Voyager difference	Significance	Effect size
District 1, school pair 1	(N=40)	(N=22)			
CTOPP-Elision	5.78	4.18	1.60	0.04	0.56
CTOPP-Blending Words	7.80	5.41	2.39	0.01	0.70
CTOPP-Blending Nonwords	3.85	2.14	1.71	0.02	0.73
CTOPP-Segmenting Words	6.85	4.05	2.80	0.02	0.63
Woodcock Word Identification	33.28	37.23	-3.95	0.32	-0.27
Woodcock Word Attack	13.00	10.05	2.95	0.26	0.31
District 1, school pair 2	(N=26)	(N=31)			
CTOPP-Elision	7.00	5.48	1.52	0.08	0.47
CTOPP-Blending Words	9.19	6.52	2.67	0.01	0.68
CTOPP-Blending Nonwords	6.12	2.29	3.83	0.00	1.42
CTOPP-Segmenting Word	10.19	4.94	5.25	0.00	1.26
Woodcock Word Identification	47.54	32.10	15.44	0.00	1.44
Woodcock Word Attack	25.00	9.97	15.03	0.00	1.76
District 2, school pair 3	(N=29)	(N=37)			
CTOPP-Elision	4.83	4.76	0.07	0.92	0.03
CTOPP-Blending Words	9.75	5.75	4.00	0.00	1.47
CTOPP-Blending Nonwords	5.67	2.41	3.26	0.00	1.11
CTOPP-Segmenting Words	8.63	3.08	5.55	0.00	1.56
Woodcock Word Identification	40.90	29.86	11.04	0.00	0.96
Woodcock Word Attack	18.69	5.08	13.61	0.00	2.06
District 2, school pair 4	(N=42)	(N=28)			
CTOPP-Elision	5.74	3.11	2.63	0.00	0.92
CTOPP-Blending Words	9.18	4.43	4.75	0.00	1.24
CTOPP-Blending Nonwords	4.73	1.93	2.80	0.00	1.12
CTOPP-Segmenting Words	7.28	2.64	4.64	0.00	1.37
Woodcock Word Identification	28.93	24.29	4.64	0.17	0.34
Woodcock Word Attack	12.10	2.68	9.42	0.00	1.42

Table 9 presents the comparative gains between Voyager and non-Voyager students during three assessment periods from years 1 and 2, based on the results from 255 students with assessment results from all three periods. Figures 1-6 show graphically that while starting at the same level at time 1, Voyager students have a higher growth rate than non-Voyager students in all six assessments.

Table 9.—Comparison of raw scores between Voyager and Non-Voyager students (T1 to T3)

Test instrument	T1		T2		T3	
	Voyager	Non-Voyager	Voyager	Non-Voyager	Voyager	Non-Voyager
CTOPP-Elision	1.12	0.92	3.96	2.68	6.92	5.36
CTOPP-Blending Words	1.48	0.84	5.30	3.07	10.34	6.42
CTOPP-Blending Nonwords	0.63	0.36	2.93	1.37	5.56	2.57
CTOPP-Segmenting Words	0.29	0.18	3.97	1.31	8.18	3.82
Woodcock Word Identification	0.89	1.04	10.72	8.14	37.15	31.54
Woodcock Word Attack	0.09	0.13	5.04	1.64	16.29	6.85

Figure 1. Comparison of growth in CTOPP-Elision

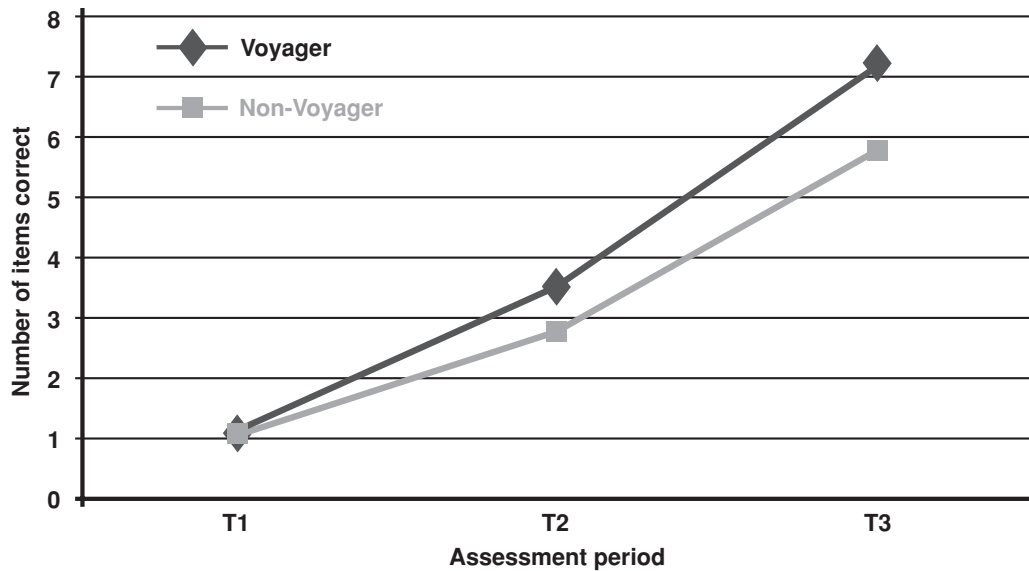


Figure 2. Comparison of growth in CTOPP-Blending Words

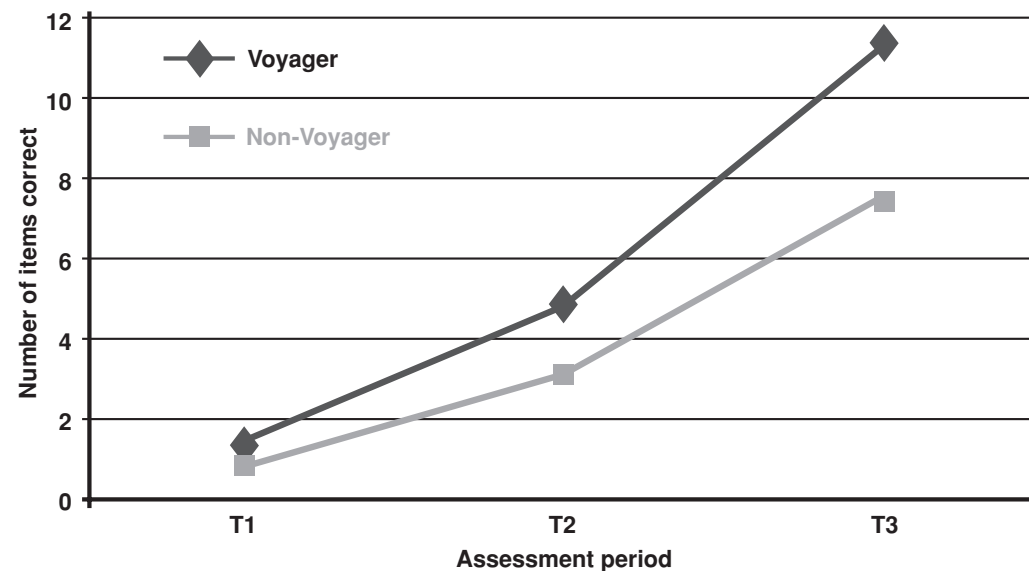


Figure 3. Comparison of growth in CTOPP-Blending Non-words

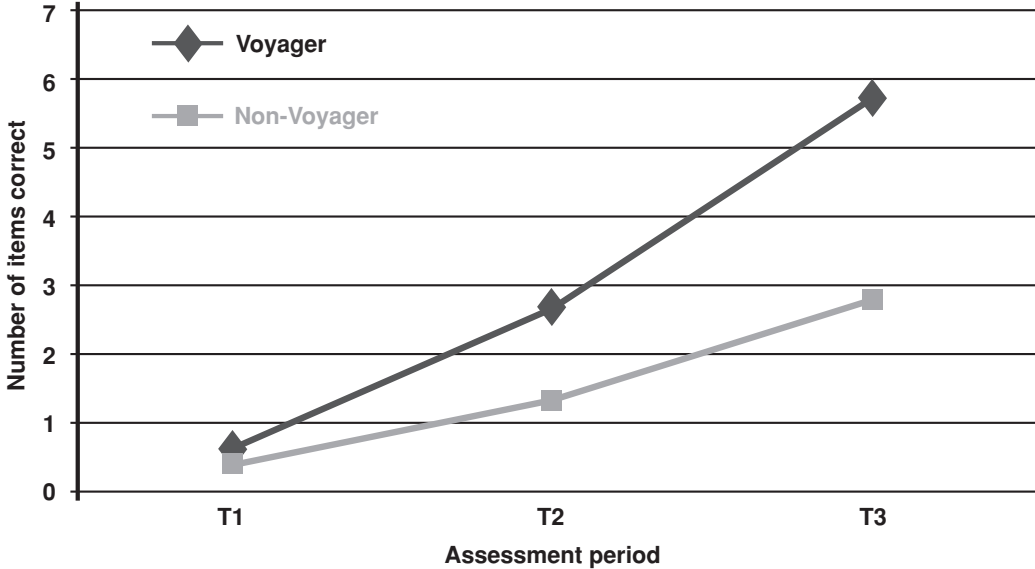


Figure 4. Comparison of growth in CTOPP-Segmenting Words

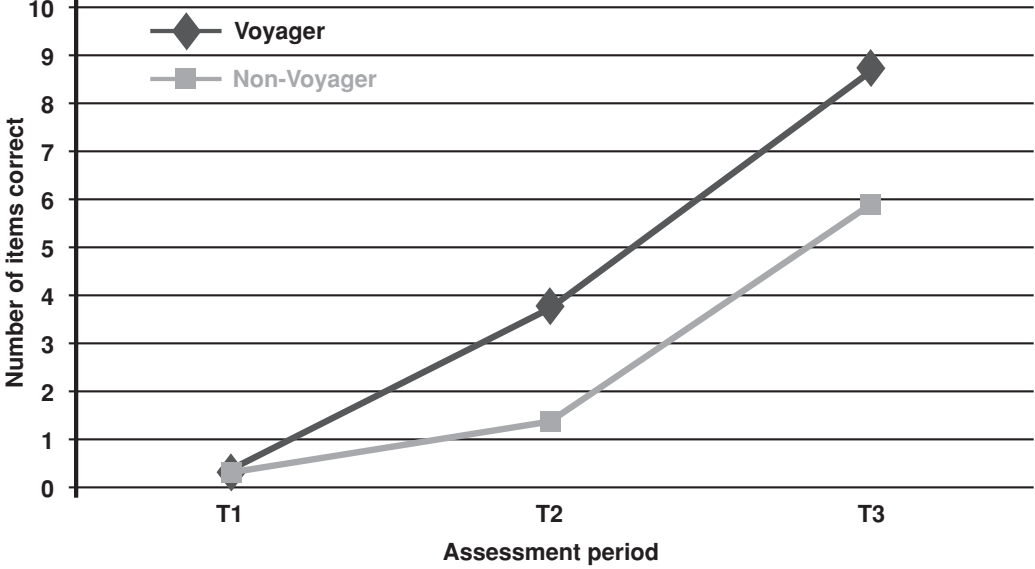


Figure 5. Comparison of growth in Woodcock Word Identification

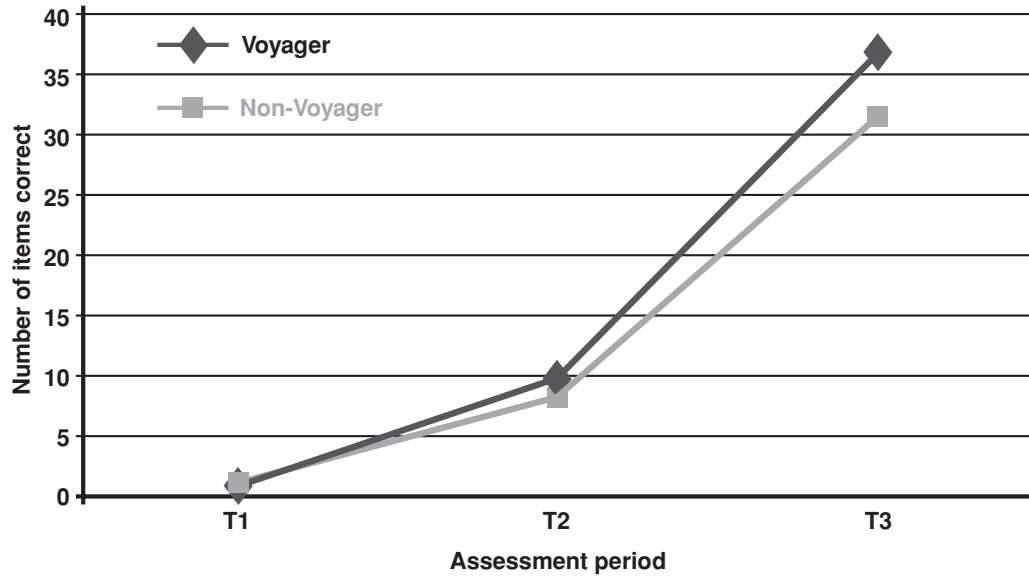
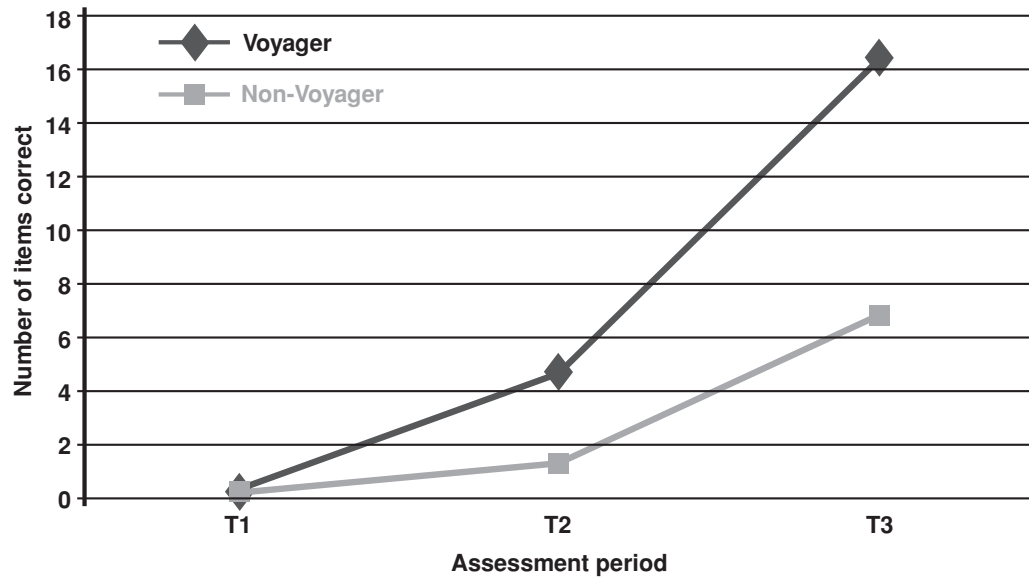


Figure 6. Comparison of growth in Woodcock Word Attack



1-3. *To what extent can the gains be attributed to the program?*

Year 2 data allow for a more extended and rigorous exploration of whether or not the observed gains can be attributed to the Voyager program. This is done using a statistically technique called HLM. Compared with other analytical models such as multiple regression and ANOVA, HLM models student achievement over time in hierarchical structure in the school setting, correcting for aggregation bias, misestimated precision, and the unit of analysis problem, thus producing more accurate results (Raudenbush and Bryk, 2002).

Model 1: Program effect (HLM)

To analyze the data, we used a three-level HLM in which time is nested within students, and students within schools. Student assessment scores were used as the dependent variable and the program status (i.e., treatment, comparison) as the main independent variable. Factors such as teacher and classroom characteristics were introduced as time-invariant covariates.⁵

At level 1, we modeled student achievement as a function of time. P₀ is the average score of the student in three time periods, and P₁ is the average growth rate of that student.

$$Y(\text{SCORE})=P_0+P_1X_1(\text{TIME})+E$$

SCORE: assessment score
TIME: testing period

Level 2 models the time-invariant variable at the student level. Teacher and classroom characteristics were disaggregated at this level.

$$P_0=B_{00}+R_0$$

$$P_1=B_{10}+B_{11}X_{11}(\text{MALE})+B_{12}X_{12}(\text{FRL})+B_{13}X_{13}(\text{TRANS})+B_{14}X_{14}(\text{TEXP})+B_{15}X_{15}(\text{CSIZE})+R_1$$

MALE: proportion of male students in the class
FRL: proportion of the students receiving free and reduced-price lunch in the class
TRANS: proportion of the students transferring out of the class
TEXP: teachers' years of experience
CSIZE: number of student in the class

⁵ Because classroom characteristics data were only collected at T2 and T3, and not T1, we could not treat them as time-varying covariates at level 1. Instead, we modeled them as time-invariant covariates at level 2 by average the values at two time periods to represent average classroom environment each student had.

Level 3 models the time-invariant variables at school level. G_{000} is the average score for students in the school from three time periods, and G_{100} is the average growth rate for students in the school. G_{101} measures the program effect on achievement gain. Because treatment schools were matched with comparison schools at the school level, the program status variable was modeled at level 3.

$$\begin{aligned}
 B_{00} &= G_{000} + U_{00} \\
 B_{10} &= G_{100} + G_{101}X_{101}(\text{VOYAGER}) + U_{10} \\
 B_{11} &= G_{110} \\
 B_{12} &= G_{120} \\
 B_{13} &= G_{130} \\
 B_{14} &= G_{140} \\
 B_{15} &= G_{150}
 \end{aligned}$$

VOYAGER: experiment status (treatment or comparison)

Table 10 presents a summary of the results for the analyses regarding program effects (detailed results are provided in appendix A). The results suggest that the Voyager program has statistically significant positive impact on student achievement gains in five out of six assessments. Only the CTOPP-Elision failed to show significant program effects, although the sign is positive. For example, for CTOPP-Blending Words (CTP-BW), on average, students in all schools had 1.59 more items correct than the previous assessment. In addition, Voyager students have 1.02 more items correct than non-Voyager students.

Table 10.—HLM results for the program effect

Variable	CTP-ELS	CTP-BW	CTP-BNW	CTP-SW	Woodcock WI	Woodcock WA
Intercept (G000)	0.94***	3.64***	1.96***	0.09**	-1.83***	-0.77**
Time (G100)	2.44***	1.59***	0.46**	1.86***	14.57***	3.55***
Voyager (G101)	0.30	1.02***	0.98***	2.11***	4.60**	4.72***
MALE (G110)	-2.81	-1.20	-3.38**	-4.20*	-4.13	-0.80
FRL (G120)	-2.56***	-0.35	-1.54***	-3.15***	-7.43*	-4.79*
TRANS (G130)	-2.34*	-1.60	-0.10	1.38	-3.39	-2.80
TEXP (G140)	-0.01	-0.01	-0.01	0.04	0.04	0.04
CSIZE (G150)	0.09**	0.02	0.04	0.04	-0.31	-0.10

*($P \leq 0.1$), **($P \leq 0.05$), ***($P \leq 0.01$). Unstandardized coefficients are presented and it is incorrect to compare them across variables.

Results from three new assessments were only collected at T3. Therefore, we use the OLS regression model with scores from DIBELS-Letter Naming Fluency from T2 as a covariate.

Model 2: Program Effect (OLS)

$$Y = P_0 + P_1X_1(\text{VOYAGER}) + P_2X_2(\text{DBL-LNF}) + P_3X_3(\text{MALE}) + P_4X_4(\text{FRL}) + P_5X_5(\text{TRANS}) + P_6X_6(\text{TEXP}) + P_7X_7(\text{CSIZE}) + U$$

Table 11.—OLS results for the program effect

Variable	DBL-ORF	WRAT3-NLW	WRAT3-SPL
Intercept	54.83***	14.24***	7.70***
Voyager	15.94***	-0.02	2.00***
DBL-LNF	0.78***	0.02***	0.07***
MALE	-49.05***	-0.48	-5.44***
FRL	-20.50***	0.00	-2.89***
TRANS	-8.03	-0.48	-0.61
TEXP	0.13	0.02*	-0.02
CSIZE	-0.60***	-0.00	-0.06*

*(P≤0.1), **(P≤0.05), ***((P≤0.01). Unstandardized coefficients are presented and it is incorrect to compare them across variables.

The results in table 11 suggest that the Voyager has statistically significant positive impacts on student achievement in two out of three new assessments. Only on WRAT3-Name/Letter Writing did the program fail to show any significant effect. The interpretation of the results is similar to that for table 10, except OLS does not model achievement gains. For instance, for DIBELS-Oral Reading Fluency (DBL-ORF), students in Voyager schools had 15.94 more items correct than non-Voyager students. In addition, student performance is highly associated with previous performance on DIBELS-Letter Naming Fluency (DBL-LNF). A correct answer to 1 item in DBL-LNF at T2 is related to a correct response of 0.78 in DBL-ORF.

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

Overall, the sample is rather homogenous, which does not allow us to detect program effect for students from different backgrounds. However, we modeled and controlled for the effects of student and classroom background such as proportion of male in the classroom, proportion of the students receiving free and reduced-price lunch, proportion of students transferring out of the classroom, and teachers’ years of experience and class size. Other variables (i.e., teacher and student attendance rate, proportion of students in different races in the classroom, proportion of students who are limited English proficient, and proportion of students with independent education plan) display little variances in the sample, and we decided to exclude them in the model. Table 12 presents descriptive statistics for the classroom data collected from teacher survey in spring 2004.

Voyager and Non-Voyager classrooms were similar in many aspects except that Voyager classes had significantly more students and higher percent of male students.

Table 12.—Descriptive statistics for classroom characteristics (T3)

Characteristics	Voyager Classes		Non-Voyager Classes		Significance
	Mean	Standard Deviation	Mean	Standard Deviation	
Total number of students	25.44	8.89	17.28	3.44	0.01
Percent male	57	12	44	11	0.02
Percent African American	97	6	96	5	0.79
Percent receiving free/reduced priced lunch	83	24	85	22	0.84
Percent limited English proficient	0	0	1	2	0.37
Percent with IEPs	7	5	11	14	0.40
Percent transferred out of classroom during the school year	17	13	29	17	0.12
Number of years teacher has taught at grade 1	9.44	8.41	7.55	6.12	0.57

As shown in tables 10 and 11, the percentage students receiving free and reduced price lunch attenuated the effects; classrooms with larger percentages of students receiving free and reduced price lunch made smaller gains in seven out of nine assessments. On CTOPP-Blending Nonwords, for example, each percentage increase in such students in the classroom is associated with 0.0154 fewer item gains (Table 10)⁶. The percentage of male students in the classroom also shows smaller gains associated with student scores in four assessments. The effects from other variables such as percent of students transferring out, teacher experience, and class size were largely not significant.

3. How does the level of implementation affect outcomes?

During the course of the program, each teacher was assessed by the Voyager program staff during regular site visits. The average final implementation score observed in May 2004 for teachers is 8.8 out of a possible 12 points⁷. The score distribution in table 13 shows that 6 of the 12 teachers demonstrate high levels of implementation, 4 had medium level of implementation, and inadequate implementation fidelity was found for 2 teachers.

⁶ The interpretation translates the proportion to percentage by dividing the coefficient with 100.

⁷ It should be noted that scoring of the IFI requires that students be given 0 if a scale component was, for some reason, not observed. This convention probably adds “noise” to the system and may deflate the relationship. A thorough examination of implementation effect usually requires using results from repeated observations.

Table 13.—Distribution of implementation scores for teachers (T3)

Implementation level	Number	Percent
High (10-12)	6	50
Medium (7-9)	4	33
Inadequate (less than 7)	2	17

To study the effect of implementation fidelity, we took an approach similar to that for the program effect by focusing on the treatment group. Model 3 is quite similar to model 1, except IFI scores (IMPLEMENT) were treated as time-invariant covariate by averaging the scores from three time periods, and modeled at level 2 to be consistent with model 1.⁷

Model 3: Implementation Effect (HLM)

Level 1

$$Y(\text{SCORE})=P_0+P_1X_1(\text{TIME})+E$$

Level 2

$$P_0=B_{00}+R_0$$

$$P_1=B_{10}+B_{11}X_{11}(\text{IMPLEMENT})+B_{12}X_{12}(\text{MALE})+B_{13}X_{13}(\text{FRL})+B_{14}X_{14}(\text{TRANS})$$

$$+B_{15}X_{15}(\text{TEXP})+B_{16}X_{16}(\text{CSIZE})+R_1$$

Level 3

$$B_{00}=G_{00}+U_{00}$$

$$B_{10}=G_{100}$$

$$B_{11}=G_{110}$$

$$B_{12}=G_{120}$$

$$B_{13}=G_{130}$$

$$B_{14}=G_{140}$$

$$B_{15}=G_{150}$$

$$B_{16}=G_{160}$$

Table 14 presents the results for the implementation model (detailed results are provided in appendix B). They suggest that implementation fidelity, measured by IFI scores, has statistically significant effect on five out of six assessments. For instance, one point of increase on IFI’s 0-12 scale is associated with 2.38 more item gains on Woodcock Word Identification.

Table 14.—HLM results for the implementation effect

Variable	CTP-ELS	CTP-BW	CTP-BNW	CTP-SW	Woodcock WI	Woodcock WA
Intercept (G000)	1.16***	4.78***	2.82**	0.14*	-1.91*	-0.84
Time (G100)	2.94***	2.59***	1.21***	3.93***	17.62***	7.64***
Voyager (G101)	0.24**	0.28***	0.11	0.47***	2.38***	1.68***
MALE (G110)	3.87	2.57	-1.15	-19.04***	-53.07*	-39.04***
FRL (G120)	-4.08***	-0.19	-1.27	-4.32***	-15.64***	-9.21***
TRANS (G130)	-3.04	-3.58	-1.15	-3.58	-6.26	3.53
TEXP (G140)	0.03	-0.02	-0.01	-0.06	-0.38*	-0.28**
CSIZE (G150)	0.14*	0.05	0.03	0.22***	0.45*	0.41***

*(P≤0.1), **(P≤0.05), ***(P≤0.01). Unstandardized coefficients are presented and it is incorrect to compare them across variables.

For results from three new assessments, we used OLS regression models (model 4) with scores from DIEBEL-Letter Naming Fluency from T2 as covariates. The results (table 15) show a lack of relationship between IFI scores and three new assessment outcomes.

Model 4: Implementation Effect (OLS)

$$Y = P_0 + P_1X_1(\text{IMPLEMENT}) + P_2X_2(\text{DBL-LNF}) + P_3X_3(\text{MALE}) + P_4X_4(\text{FRL}) + P_5X_5(\text{TRANS}) + P_6X_6(\text{TEXP}) + P_7X_7(\text{CSIZE}) + U$$

Table 15.—OLS results for the implementation effect

Variable	DBL-ORF	WRAT3-NLW	WRAT3-SPL
Intercept	134.21***	16.50***	14.90***
IMPLEMENT	1.748	0.04	0.20
DBL-LNF	0.71***	0.00	0.05***
MALE	-205.23**	-4.00	-18.36**
FRL	-11.58	-0.55	-2.38
TRANS	39.48	-1.63	-0.60
TEXP	-1.10*	0.00	-0.09
CSIZE	-0.24	0.02	-0.01

*(P≤0.1), **(P≤0.05), ***(P≤0.01). Unstandardized coefficients are presented and it is incorrect to compare them across variables.

We also examined how student assessment outcomes are affected by implementation groupings: high (10-12), medium (7-9) and inadequate (0-6). The regression results are similar to those where we treated implementation scale as a continuous variable. Looking at the scores between students in high implementation classes and low implementation classes, we found that students in the high implementation classes had significantly larger gains than those from low implementation classes in

both Woodcock Word Identification and Word Attack, as well as had higher scores in DIBELS-Oral Reading Fluency and WRAT3-Spelling (Table 16).

Table 16.—Comparison of student raw scores between high and low implementation classes (T3)

Test Instrument	High (N=55)		Low (N=42)		Diff (H-L)	Sig
	Mean	Standard Deviation	Mean	Standard Deviation		
DIBELS-Oral Reading Fluency (status)	42.02	26.16	30.83	20.83	11.19	0.03
Woodcock Word Identification (gain)	37.98	13.41	28.93	14.97	9.05	0.00
Woodcock Word Attack (gain)	16.36	9.89	12.10	9.14	4.26	0.03
WRAT3-Spelling (status)	5.27	3.02	3.93	2.62	1.34	0.02

Methodological Exploration

In addition to providing answers to the evaluation questions, the researchers also explored a new methodological front with the Voyager data. Previous experience with using HLM methods has required at least three data points to model individual growth. We experimented with a parallel scale approach in analyzing scores at two data points using selected student achievement data in 2003. The results and discussions are presented in appendix C. The technique has the potential to apply a rigorous method on data analysis without imposing an intensive requirement on data collection.

Conclusion

The results from year 2 of the evaluation continue to support the findings from year 1 regarding the efficacy of the Voyager program. Based on individual-level data from 255 students, we found that Voyager students in two urban school districts and all four school pairs significantly outperformed their counterparts in eight out of nine reading assessments. On the measure with no significant difference, most students in both treatment and comparison schools correctly answered all of the questions in the test. In addition, Voyager students have made greater gains than non-Voyager students from 2002 to 2004 on measures used in both years. The difference in achievement progress between the two groups is large and statistically significant at the 0.05 level, with an average effect size of 0.82. Further, the average scores of the Voyager students at the end of first grade were largely at or above the national average, while those of comparison students generally remained below the national average.

We used HLM to examine the program and implementation effects on student achievement from both years, and OLS model to assess these effects for the results from three new assessments in 2004. We found that the Voyager program has statistically significant positive impacts on student achievement in seven out of nine assessments. Among classroom characteristics, the percentage of males and students with free and reduced price lunch attenuated the effects; classrooms with larger percentages of males and/or students receiving free and reduced price lunch made smaller gains. The effects from other variables such as percent of students transferring out, teachers' experience, and class size were largely not significant. Results also suggest that implementation fidelity, measured by IFI scores, has a statistically significant effect on five out of nine assessments. 8

References

- Frechtling, J., Silverstein, G., and Zhang, X. (2003). *An evaluation of the Voyager 100% Literacy System: Results from a study of kindergarten students in inner city schools*. (Prepared under contract to Voyager Expanded Learning and The Council of Great City Schools). Rockville, MD: Westat.
- Lipsey, M.W., and Wilson, D.B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.
- 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.W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2(2):173-185.
- Raudenbush, S.W., and Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.
- 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 Test - Revised*. Circle Pines, MN: American Guidance Service.
- Woodcock, R.W. (1998). *Woodcock Reading Mastery Test - Revised (Forms G and H) Examiner's Manual*. Circle Pines, MN: American Guidance Service.

Appendix A. Detailed HLM results for program effects

Table A-1: HLM results for program effect (CTOPP Elision)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, P0 INTRCPT2, B00 INTRCPT3, G000	2.086718	0.229332	9.099	7	0.000
For TIME slope, P1 INTRCPT2, B10 INTRCPT3, G100	2.953777	0.933093	3.166	6	0.022
EXP, G101	0.425729	0.344879	1.234	6	0.264
For MALE, B11 INTRCPT3, G110	-3.527971	1.992267	-1.771	248	0.077
For CLSIZE, B12 INTRCPT3, G120	0.051618	0.038496	1.341	248	0.181
For TCHEXP, B13 INTRCPT3, G130	-0.005980	0.025922	-0.231	248	0.818
For FRLUNCH, B14 INTRCPT3, G140	-2.347038	0.803913	-2.920	248	0.004
For TRANS, B15 INTRCPT3, G150	1.595122	1.596408	-0.999	248	0.319

Final estimation of level-1 and level-2 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1, R0	0.93428	0.87289	246	184.93908	>.500
TIME slope, R1	0.16577	0.02748	241	123.40942	>.500
level-1, E	3.03719	9.22451			

Final estimation of level-3 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1/INTRCPT2, U00	0.37964	0.14413	7	12.12191	0.096
TIME/INTRCPT2, U10	0.17239	0.02972	6	6.32980	0.387

Table A-2: HLM results for program effect (CTOPP Blending Words)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, P0 INTRCPT2, B00 INTRCPT3, G000	2.537880	0.482435	5.261	7	0.000
For TIME slope, P1 INTRCPT2, B10 INTRCPT3, G100	2.771134	1.686105	1.644	6	0.151
EXP, G101	1.303306	0.627329	2.078	6	0.082
For MALE, B11 INTRCPT3, G110	-2.538992	3.578152	-0.710	248	0.479
For CLSIZE, B12 INTRCPT3, G120	0.136520	0.070290	1.942	248	0.053
For TCHEXP, B13 INTRCPT3, G130	0.036397	0.046784	0.778	248	0.437
For FRLUNCH, B14 INTRCPT3, G140	-2.454864	1.467990	-1.672	248	0.095
For TRANS, B15 INTRCPT3, G150	1.361936	2.877938	0.473	248	0.636

Final estimation of level-1 and level-2 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1, R0	0.36393	0.13245	246	131.90825	>.500
TIME slope, R1	2.01338	4.05372	241	252.22008	0.297
level-1, E	4.67857	21.88900			

Final estimation of level-3 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1/INTRCPT2, U00	1.12107	1.25680	7	24.07217	0.001
TIME/INTRCPT2, U10	0.54860	0.30096	6	7.79465	0.253

Table A-3: HLM results for program effect (CTOPP Blending Nonwords)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, P0 INTRCPT2, B00 INTRCPT3, G000	1.342451	0.328279	4.089	7	0.006
For TIME slope, P1 INTRCPT2, B10 INTRCPT3, G100	2.072399	0.714868	2.899	6	0.028
EXP, G101	0.891680	0.277511	3.213	6	0.021
For MALE, B11 INTRCPT3, G110	-3.122728	1.509421	-2.069	248	0.039
For CLSIZE, B12 INTRCPT3, G120	0.028347	0.030946	0.916	248	0.361
For TCHEXP, B13 INTRCPT3, G130	0.004593	0.019102	0.240	248	0.810
For FRLUNCH, B14 INTRCPT3, G140	-1.740693	0.644850	-2.699	248	0.008
For TRANS, B15 INTRCPT3, G150	-0.499674	1.213669	-0.412	248	0.681

Final estimation of level-1 and level-2 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1, R0	0.06396	0.00409	246	173.62362	>.500
TIME slope, R1	0.08657	0.00749	241	185.21510	>.500
level-1, E	2.46826	6.09230			

Final estimation of level-3 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1/INTRCPT2, U00	0.83385	0.69530	7	40.61560	0.000
TIME/INTRCPT2, U10	0.24814	0.06157	6	9.90633	0.128

Table A-4: HLM results for program effect (CTOPP Segmenting Words)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, P0 INTRCPT2, B00 INTRCPT3, G000	1.193536	0.449274	2.657	7	0.033
For TIME slope, P1 INTRCPT2, B10 INTRCPT3, G100	3.027443	1.065239	2.842	6	0.030
EXP, G101	1.177987	0.395154	2.981	6	0.026
For MALE, B11 INTRCPT3, G110	-3.132612	2.242136	-1.397	248	0.164
For CLSIZE, B12 INTRCPT3, G120	-0.012766	0.044023	-0.290	248	0.772
For TCHEXP, B13 INTRCPT3, G130	0.054429	0.028840	1.887	248	0.060
For FRLUNCH, B14 INTRCPT3, G140	-1.852112	0.917599	-2.018	248	0.044
For TRANS, B15 INTRCPT3, G150	-0.476628	1.797134	-0.265	248	0.791

Final estimation of level-1 and level-2 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1, R0	0.11490	0.01320	246	189.27227	>.500
TIME slope, R1	0.08444	0.00713	241	169.66724	>.500
level-1, E	3.75269	14.08265			

Final estimation of level-3 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1/INTRCPT2, U00	1.10849	1.22875	7	33.96503	0.000
TIME/INTRCPT2, U10	0.48086	0.23122	6	12.45093	0.052

Table A-5: HLM results for program effect (Woodcock Word Identification)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, P0 INTRCPT2, B00 INTRCPT3, G000	2.493704	0.833285	2.993	7	0.021
For TIME slope, P1 INTRCPT2, B10 INTRCPT3, G100	16.424575	3.914141	4.196	6	0.007
EXP, G101	3.924973	1.472002	2.666	6	0.037
For MALE, B11 INTRCPT3, G110	-12.439942	8.346174	-1.490	248	0.137
For CLSIZE, B12 INTRCPT3, G120	-0.110903	0.164844	-0.673	248	0.502
For TCHEXP, B13 INTRCPT3, G130	0.053241	0.107607	0.495	248	0.621
For FRLUNCH, B14 INTRCPT3, G140	-8.705373	3.422293	-2.544	248	0.012
For TRANS, B15 INTRCPT3, G150	0.082957	6.700063	0.012	248	0.990

Final estimation of level-1 and level-2 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1, R0	0.35336	0.12486	246	96.66166	>.500
TIME slope, R1	0.24626	0.06064	241	92.40279	>.500
level-1, E	13.99899	195.97167			

Final estimation of level-3 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1/INTRCPT2, U00	0.58104	0.33761	7	6.06185	>.500
TIME/INTRCPT2, U10	0.04994	0.00249	6	4.11663	>.500

Table A-6: HLM results for program effect (Woodcock Word Attack)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, P0 INTRCPT2, B00 INTRCPT3, G000	0.817876	0.580868	1.408	7	0.202
For TIME slope, P1 INTRCPT2, B10 INTRCPT3, G100	5.578104	2.191205	2.546	6	0.007
EXP, G101	3.112034	0.925995	3.361	6	0.018
For MALE, B11 INTRCPT3, G110	-5.446796	4.646517	-1.172	248	0.243
For CLSIZE, B12 INTRCPT3, G120	-0.072327	0.103507	-0.699	248	0.485
For TCHEXP, B13 INTRCPT3, G130	0.039693	0.057698	0.688	248	0.492
For FRLUNCH, B14 INTRCPT3, G140	-6.090325	2.141579	-2.844	248	0.005
For TRANS, B15 INTRCPT3, G150	-1.782924	2.141579	-0.473	248	0.636

Final estimation of level-1 and level-2 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1, R0	0.30447	0.09270	246	80.29258	>.500
TIME slope, R1	1.46215	2.13790	241	136.87477	>.500
level-1, E	6.50942	42.37253			

Final estimation of level-3 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1/INTRCPT2, U00	1.24408	1.54774	7	16.79926	0.019
TIME/INTRCPT2, U10	0.09325	0.00870	6	4.35985	>.500

Appendix B. Detailed HLM results for implementation effects

Table B-1. HLM results for implementation model (CTOPP Elision)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, P0 INTRCPT2, B00 INTRCPT3, G000	2.512629	0.293025	8.575	3	0.000
For TIME slope, P1 INTRCPT2, B10 INTRCPT3, G100	1.640752	0.228584	7.178	332	0.000
For IMPLEMEN, B11 INTRCPT3, G110	0.277104	0.114195	2.427	332	0.016
For MALE, B12 INTRCPT3, G120	-6.099064	6.246048	-0.976	332	0.330
For CLSIZE, B13 INTRCPT3, G130	0.109256	0.058268	1.875	332	0.061
For TCHEXP, B14 INTRCPT3, G140	-0.040663	0.044985	-0.904	332	0.367
For FRLUNCH, B15 INTRCPT3, G150	-3.323957	1.084415	-3.065	332	0.003
For TRANS, B16 INTRCPT3, G160	-1.045598	2.820673	-0.371	332	0.711

Final estimation of level-1 and level-2 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1, R0	0.70963	0.50358	114	133.10142	0.107
level-1, E	3.38748	11.47500			

Final estimation of level-3 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1/INTRCPT2, U00	0.02296	0.00053	3	3.85721	0.276

Table B-2. HLM results for implementation model (CTOPP Blending Words)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, P0 INTRCPT2, B00 INTRCPT3, G000	2.895358	0.588809	4.917	3	0.011
For TIME slope, P1 INTRCPT2, B10 INTRCPT3, G100	3.120612	0.439715	7.097	332	0.000
For IMPLEMEN, B11 INTRCPT3, G110	0.028068	0.215621	0.130	332	0.897
For MALE, B12 INTRCPT3, G120	3.748319	11.744907	0.319	332	0.750
For CLSIZE, B13 INTRCPT3, G130	0.152585	0.111577	1.368	332	0.172
For TCHEXP, B14 INTRCPT3, G140	0.079452	0.083862	0.947	332	0.345
For FRLUNCH, B15 INTRCPT3, G150	-1.790202	2.091313	-0.856	332	0.393
For TRANS, B16 INTRCPT3, G160	1.494977	5.263404	0.284	332	0.777

Final estimation of level-1 and level-2 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1, R0	0.19108	0.03651	114	95.78653	>.500
level-1, E	6.52435	42.56715			

Final estimation of level-3 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1/INTRCPT2, U00	0.39414	0.15535	3	6.20428	0.101

Table B-3. HLM results for implementation model (CTOPP Blending Nonwords)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, P0 INTRCPT2, B00 INTRCPT3, G000	1.852796	0.399801	4.634	3	0.021
For TIME slope, P1 INTRCPT2, B10 INTRCPT3, G100	1.244123	0.185942	6.691	332	0.000
For IMPLEMEN, B11 INTRCPT3, G110	0.110453	0.095784	1.153	332	0.250
For MALE, B12 INTRCPT3, G120	-5.579502	5.142061	-1.085	332	0.279
For CLSIZE, B13 INTRCPT3, G130	0.029759	0.052223	0.570	332	0.569
For TCHEXP, B14 INTRCPT3, G140	-0.029104	0.035743	-0.814	332	0.416
For FRLUNCH, B15 INTRCPT3, G150	-1.464949	1.060185	-1.382	332	0.168
For TRANS, B16 INTRCPT3, G160	-1.824042	2.249907	-0.811	332	0.418

Final estimation of level-1 and level-2 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1, R0	0.05022	0.00252	114	78.60368	>.500
level-1, E	2.75871	7.61049			

Final estimation of level-3 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1/INTRCPT2, U00	0.63203	0.39947	3	21.78993	0.000

Table B-4. HLM results for implementation model (CTOPP Segmenting Words)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, P0 INTRCPT2, B00 INTRCPT3, G000	1.922128	0.489101	3.930	3	0.058
For TIME slope, P1 INTRCPT2, B10 INTRCPT3, G100	2.379934	0.279790	8.506	332	0.000
For IMPLEMEN, B11 INTRCPT3, G110	0.079421	0.142131	0.559	332	0.576
For MALE, B12 INTRCPT3, G120	-1.032800	7.663339	-0.135	332	0.893
For CLSIZE, B13 INTRCPT3, G130	-0.022895	0.076279	-0.300	332	0.764
For TCHEXP, B14 INTRCPT3, G140	0.069668	0.053656	1.298	332	0.195
For FRLUNCH, B15 INTRCPT3, G150	-1.576803	1.496472	-1.054	332	0.293
For TRANS, B16 INTRCPT3, G160	-0.866959	3.374331	-0.257	332	0.797

Final estimation of level-1 and level-2 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1, R0	0.08378	0.00702	114	81.38314	>.500
level-1, E	4.15119	17.23241			

Final estimation of level-3 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1/INTRCPT2, U00	0.65580	0.43007	3	12.99994	0.005

Table B-5. HLM results for implementation model (Woodcock Word Identification)

Fixed Effect		Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, INTRCPT2, INTRCPT3,	P0 B00 G000	2.740877	1.203277	2.278	3	0.097
For TIME slope, INTRCPT2, INTRCPT3,	P1 B10 G100	13.388284	0.962139	13.915	332	0.000
For IMPLEMEN, INTRCPT3,	B11 G110	1.613706	0.464582	3.473	332	0.001
For MALE, INTRCPT3,	B12 G120	-47.957553	25.414018	-1.887	332	0.060
For CLSIZE, INTRCPT3,	B13 G130	0.094823	0.236995	0.400	332	0.689
For TCHEXP, INTRCPT3,	B14 G140	-0.300345	0.183016	-1.641	332	0.101
For FRLUNCH, INTRCPT3,	B15 G150	-9.314318	4.409157	-2.112	332	0.035
For TRANS, INTRCPT3,	B16 G160	-2.441314	11.478436	-0.213	332	0.832

Final estimation of level-1 and level-2 variance components:

Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1, level-1,	R0 E	0.32548 14.27646	0.10594 203.81741	114	79.63225	>.500

Final estimation of level-3 variance components:

Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1/INTRCPT2,	U00	0.04877	0.00238	3	2.90796	>.500

Table B-6. HLM results for implementation model (Woodcock Word Attack)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, P0 INTRCPT2, B00 INTRCPT3, G000	1.675100	0.648073	2.585	3	0.073
For TIME slope, P1 INTRCPT2, B10 INTRCPT3, G100	5.366916	0.517887	10.363	332	0.000
For IMPLEMEN, B11 INTRCPT3, G110	1.106425	0.250116	4.424	332	0.000
For MALE, B12 INTRCPT3, G120	3.748319	13.681528	-2.144	332	0.033
For CLSIZE, B13 INTRCPT3, G130	0.132221	0.127607	1.036	332	0.301
For TCHEXP, B14 INTRCPT3, G140	-0.120289	0.098518	-1.221	332	0.223
For FRLUNCH, B15 INTRCPT3, G150	-6.637822	2.374170	-2.796	332	0.006
For TRANS, B16 INTRCPT3, G160	-5.677368	6.178910	-0.919	332	0.359

Final estimation of level-1 and level-2 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1, R0	0.18382	0.03379	114	84.84657	>.500
level-1, E	7.68453	59.05207			

Final estimation of level-3 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1/INTRCPT2, U00	0.04837	0.00234	3	4.14594	0.245

Appendix C. Using HLM to Analyze Pre-Test and Post-Test Scores

The advantages of using multilevel model in school setting where the observations are not independent are well documented (Raudenbush and Bryk, 2002). However, HLM requires at least three data points for each student. In this section, we explore using HLM to analyze achievement data in two data points using parallel scale approach. This approach requires collection of item-level data from an assessment with a relatively large number of items.

Model

We used a three-level HLM model.⁹ Level 1 models the relationship between two time points (pre-test and post-test) for an individual student on a given instrument. Level 2 approximates the individual effects on scores. Level 3 models the classroom and school effects on scores.

At level 1, TIME identifies whether the score is from pre-test (0) or post-test (1). Therefore, the intercept (P0) represents the expected pre-test score of an individual, and the slope (P1) represents the expected gain score of that individual. To fit a regression line for each relationship, conventional wisdom calls for at least three points of data. However, we only have two data points for each relationship—pre-test and post-test. Thus, we used a split-half technique to establish two parallel scales by first creating pairs of items matched on their standard deviations and then randomly assigning one item from each matched pairs to Scale A or B (Lyons et al., 2002). For example, the raw scores from 64 items in Woodcock Word Attack were randomly split and assigned into scale A and B, each of which contains 32 subscales, for both pre-test and post-test respectively. Consequently, each person has four eligible data points (two pre-scores and two post scores).

Level 1

$$Y(\text{SCORE})=P_0+P_1X_1(\text{TIME})+E$$

Level 2 models the individual effects on the scores. The intercept denotes the average pre-test scores for all individuals within a class and the slope represents the average gain.

Level 2

$$P0=B_{00}+R_0$$

$$P1=B_{10}+R_1$$

⁹ The model assumes linearity and normality. We also tried to use generalized hierarchical linear models to approximate the nonlinear structural models and non-normally distributed errors, treating scores as count data. We first used a transformation $Y_j=\log(1+Y_j)$ to eliminate the out-of-bounds predictors. Then, we used a three-level nonlinear model for estimation. The results somewhat are similar to those from the current model.

Level 3 models the classroom effects on scores. G_{000} is the grand mean of the pre-test score, while G_{100} is the grand mean of gain score. A class' average pre-test score can be predicted by the grand mean (G_{000}) plus the random effect from the classroom (U_{00}). Similarly, the class' gain score can be predicted by the grand mean (G_{100}) plus the random effect from the classroom (U_{100}). The experiment status as well as other classroom and teacher characteristics variables is modeled at this level.

Level 3

$$B_{00} = G_{000} + U_{00}$$

$$B_{10} = G_{100} + G_{101}X_{101}(\text{VOYAGER}) + G_{102}X_{102}(\text{MALE}) + B_{103}X_{103}(\text{FRL}) + B_{104}X_{104}(\text{TRANS}) \\ + B_{105}X_{105}(\text{TEXP}) + B_{106}X_{106}(\text{CSIZE}) + U_{10}$$

Results

Tables D-1 and D-2 present the results for Woodcock Word Identification and Woodcock Word Attack. Take D-1 for example—at level 1, the reliability estimate for P1 is 0.94, suggesting that 94 percent of the true variance was contained in the observation. At level 2, the reliability estimate for B10 is 0.56. The sign for covariance between B00 and B10 is positive, which suggests that classes with higher pre-test scores are likely to have larger post-test gains.

Because the evaluation is about how changes in the independent variable can be expected to affect the overall population mean, we chose the population-average model (instead of unit-specific model). In this model, the mean and gain scores for an individual were predicted only by their grand means. The mean coefficient is 0.48, which was then multiplied by 2 to account for the parallel split. The resulting coefficient is 0.96, meaning that the expected grand mean for pre-test score is 0.96 out of a test of 64 items. The coefficient for gain is 3.69, which became 7.38 after transformation. This means that the expected gain for a student is 7.38 points. Both estimates are statistically significant. However, none of the other variables are particularly significant, except for teacher experience.

For Woodcock Word Attack, the results suggest that the expected grand mean for pre-test score is 0.22, while the expected gain for a student is 1.26. However, Voyager students on average gained 3.02 points more than their counterparts in the comparison group, and the difference was significant. Teacher experience also has statistically positive effect while other variable do not appear to have significant effect on scores.

Table C-1. HLM results for program model using parallel scale technique (Woodcock Word Identification)

Fixed Effect		Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, INTRCPT2, INTRCPT3,	P0 B00 G000	0.483526	0.149260	3.239	16	0.006
For TIME slope, INTRCPT2, INTRCPT3,	P1 B10 G100	3.690924	0.366100	10.082	11	0.000
CLSIZE_1,	G101	0.027034	0.077679	0.348	11	0.734
FRLUNC_1,	G102	-1.981142	3.201514	-0.619	11	0.548
EXP_1,	G103	0.923888	0.655172	1.410	11	0.186
GENDER_1,	G104	-4.186124	6.263486	-0.668	11	0.517
TCHEXP_1,	G105	0.063720	0.034736	1.834	11	0.093

Final estimation of level-1 and level-2 variance components:

Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1,	R0	2.40389	5.77870	255	3617.06240	0.000
TIME slope,	R1	3.81856	14.58137	255	2067.93083	0.000
level-1,	E	0.94810	0.89889			

Final estimation of level-3 variance components:

Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1/INTRCPT2,	U00	0.14477	0.02096	16	15.43014	>.500
TIME/INTRCPT2,	U10	1.14172	1.30352	11	38.52393	0.000

Table C-2. HLM results for program effect using parallel scale technique (Woodcock Word Attack)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, P0 INTRCPT2, B00 INTRCPT3, G000	0.112809	0.075085	1.502	14	0.155
For TIME slope, P1 INTRCPT2, B10 INTRCPT3, G100	0.630532	0.218334	10.082	9	0.019
CLSIZE_1, G101	-0.041485	0.050772	-0.817	9	0.435
FRLUNC_1, G102	1.595730	1.166314	1.368	9	0.205
EXP_1, G103	1.507299	0.399332	3.775	9	0.005
GENDER_1, G104	1.643147	2.290386	0.717	9	0.491
TCHEXP_1, G105	0.053966	0.020457	2.638	9	0.027

Final estimation of level-1 and level-2 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1, R0	1.17773	1.38705	230	3256.33522	0.000
TIME slope, R1	1.93795	3.75566	230	1842.97605	0.000
level-1, E	0.45288	0.20510			

Final estimation of level-3 variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1/INTRCPT2, U00	0.10122	0.01025	14	13.20679	>.500
TIME/INTRCPT2, U10	0.60790	0.36955	9	37.25511	0.000

While the OLS and HLM produced somewhat similar results in the parameter estimates (HLM can predict growth in addition to status), the results on the significant test varied (table D-3). In general, HLM produces more conservative results than OLS. By modeling the nesting structure, it produces a more accurate estimate of standard error. In fact, the standard errors of the parameter estimates from the OLS tend to be underestimated. As a result, HLM is less likely than OLS to result in statistically significant findings. In addition, HLM models place more stringent requirement on the completeness of data. Cases which contain one missing variable will be automatically excluded. Therefore, HLM model analyzed a smaller sample than that for the OLS. Also, it is noteworthy that the HLM results are based on a very small number of level 3 units, which may jeopardize the reliability of the findings.

Table C-3. Comparison of OLS and HLM results

Variable	OLS		HLM	
	Coefficient	P	Coefficient	P
Woodcock Word Identification				
Grand mean pre-score	2.00	0.02	0.96	0.01
Grand mean gain score	NA	NA	7.38	0.00
Voyager	3.94	0.04	1.84	0.19
Percent male	-0.06	0.11	-8.38	0.52
Percent free lunch	0.00	0.15	-3.96	0.55
Teacher experience	0.11	0.16	0.12	0.09
Class size	0.05	0.49	0.06	0.73
Woodcock Word Attack				
Grand mean pre-score	14.71	0.54	0.22	0.16
Grand mean gain score	NA	NA	1.26	0.02
Voyager	2.30	0.00	3.02	0.01
Percent male	-0.12	0.16	3.28	0.49
Percent free lunch	-0.05	0.94	3.20	0.21
Teacher experience	0.11	0.01	0.10	0.03
Class size	0.09	0.40	-0.08	0.44



1650 Research Boulevard
Rockville, Maryland 20850
(301) 251-1500