Evaluating Effect of Reflex® on Math Fact Fluency in Grades 2 & 3

David I. Rudel*

March 13, 2017

1 Study Characteristics

1.1 Intervention Condition

Reflex is an online, game-based system for developing math fact fluency in schoolchildren. It is provided by ExploreLearning, a division of Cambium-Learning. Reflex maintains an internal student model to facilitate adaptive instruction and individualized practice on math facts. It uses a fact-family approach, teaching groups of related facts together. For example, a student may receive coaching on 2+6, 6+2, 8-2, and 8-6 on the same day. A student's daily work in Reflex generally comprises 4 phases:

1. An assessment component monitoring progress posed in a game environment that minimizes distraction

2. A coaching session where the student learns a new set of related facts or receives remedial work on a previously learned set

3. A practice game combining newly learned facts with facts the student is developing

4. Intense practice under time pressure on facts the student has demonstrated at least partial fluency

*Senior Principal Data Scientist—ExploreLearning

The assessment component uses a combination of several games, some of which present facts aligned vertically while others present facts aligned horizontally. The coaching session uses a cover-copy-compare strategy to introduce facts followed by a fill-in-the-blank session where the student completes an open fact sentence with one or two missing terms. The third component uses horizontally aligned facts and provides interactive feedback to missed facts. The intense practice component differs from the rest in that the student is given multiple facts and chooses one to answer. This choice provides agency to the student, as it affects outcomes in the game (e.g., the fact chosen determines which direction an on-screen character moves).

Reflex has individualized practice recommendations. The median total time in the system for second and third graders to complete these recommendations is 15-16 minutes per day, with earlier days generally taking longer than later ones. Students do not always meet the daily practice target due to lack of time or limited technological resources. Once the recommended practice is complete for a day, an on-screen indicator illuminates, and the student is allowed to spend time on non-practice, motivational aspects of the system, such as using tokens to buy new clothes for his avatar.

Reflex has been sold commercially since 2011. It is delivered on an annual subscription basis to thousands of schools. A time-limited free trial is available, and interested teachers can apply for grants providing free access for one year. Subscriptions are sold at teacher-, site-, and district-wide levels.

Participating teachers assigned to the intervention condition undertook a standard, 90-minute training webinar acquainting them with the system and best practices. Approximately 50% of all new Reflex subscriptions included such training in spring 2016.

Students use Reflex directly; no teacher involvement occurs within a Reflex session. Teachers support students indirectly by encouraging students and cultivating their enthusiasm, including the distribution of milestone certificates provided by the system. Teachers also, of course, need to schedule time for students to play Reflex and supervise student usage. Reflex provides teachers reports showing progress and usage of each student.

Reflex provides three options for the pool of facts a student learns:

- Addition and Subtraction 0-10: Addition facts whose terms are within 0-10 and their associated subtraction facts

- Multiplication and Division 0-10: Multiplication facts whose factors are in the range 0-10 and their associated division facts

- Multiplication and Division 0-12: Multiplication facts whose factors are in the range 0-12 and their associated division facts

Students assigned to the intervention condition began in the addition / subtraction assignment if they were in second grade and in the multiplication / division 0-10 assignment if they were in third grade. Teachers had the ability to switch students on an individual basis to other assignments at their own discretion. Sixteen of the 37 second grader using Reflex were switched into multiplication/division before the posttest. Thus, some of their time spent in Reflex was dedicated to above-grade-level items that were not on the posttest.

The recommended usage for Reflex is 3 days per week. The four teachers achieved weekly usages of 2.6, 3.3, 3.4, and 1.5. These values include all days on which a login was made, even if the student was practicing facts outside the range of testing.

The average usage across all students was 2.7 days/week.

The median time spent in Reflex during the study's was 59 minutes a week, which includes time spent in non-instructional aspects of the system such as browsing an in-product store to buy virtual items using tokens earned in games or cases where a student logged in from home and forgot to log off.

Reflex requires individual accounts with individual passwords. A user in the comparison group could only have used Reflex by logging into the account of another student.

Post-survey questionnaires were given to all teachers. Two teachers from the intervention condition returned questionnaires, both indicating they relied on Reflex as their primary means of developing math fact fluency during the course of the study.

1.2 Comparison Condition

This study used a business as usual comparison condition. Math fluency in general and math fact fluency in particular are required by the *Florida Math Standards* and *Common Core State Standards* for grades 2 and 3. *Florida Math Standard* MAFS.2.OA.2.2 and *Common Core State Standard* 2.MD.2, have identical wordings: "Fluently add and subtract within 20 using mental strategies. By end of Grade 2, know from memory all sums of two one-digit numbers." Similarly, *Florida Math Standard* MAFS.OA.C.7 and *Common Core State Standard*3.OA.C.7 read "Fluently multiply and divide within 100, using strategies such as the relationship between multiplication and division (e.g., knowing that 8 5 = 40, one knows 40 5 = 8) or properties of operations. By the end of Grade 3, know from memory all products of two one-digit numbers."

Additionally, *Common Core State Standards* specify a number of general computational fluency requirements for which facility with math facts are foundational (Standards 2.NBT.B.5, 2.NBT.B.6, 2.NBT.B.7, 3.OA.A., 3.NBT.A.2). Florida's standards retain these requirements.

Post-survey questionnaires were given to all teachers. Teachers in the comparison condition were asked to describe methods they used to develop math fact fluency and the time they spent on this goal. Two of the four teachers in the comparison condition returned these questionnaires. Their comments are provided below verbatim. We have also included data on the average fluency gain for each comparison class, including the two that did not return questionnaires.

The survey asked teachers how many hours a month were spent on developing math fact fluency. One teacher specified her answer in terms of minutes per day. The wrote "20 hours" in the blank.

Table 1: Post-Study Comparison Group Responses

Grade	Average Gain	Strategies	Time Spent (Hours per month)
3	0.88	(Did not return survey)	N/A
3	0.81	flash cards, timed tests, repetition, math fact raps	20 hours
2	0.53	(Did not return survey)	N/A
2	-0.01	ten marks, flash cards, fast facts, center work	(time everyday) 10 minutes

Given the average fluency gains, we surmise that the other two teachers likely spent considerably more than 10 minutes a day on math fact fluency. The grade 3 responder had a group of high-achieving students, so it is possible homework was assigned on math fact fluency, as it is hard to imagine that 20 hours of class time a month was spent on the topic.

1.3 Setting

Teachers from a Florida school in a metropolitan area participated in this study. The demographic data provided by the school indicate it is a majority-minority school. 57% of its second- and third-grade students are Hispanic or Latino, and 31% are Caucasian. The data provided indicate that 28% have low English proficiency and 17% are on free or reduced lunch.

1.4 Participants

The participating students are generally demographically similar to the full population of second- and third-grade students in terms of exceptional student status, race, gender, and economic status. In all cases we relied on information received from the school.

2 Study Design and Analysis

2.1 Sample Formation

The school was identified by project personnel owing to its previous interest in Reflex. The school was offered a discount on a later subscription in exchange for participation. After logistical discussions to ensure that the school had sufficient technical resources to allow usage of a computer-delivered intervention, teachers were asked to volunteer for participation. Nine teachers initially volunteered to have their homeroom students take part. One of these homeroom classes was taught by another teacher who also taught her own homeroom, so the 9 classes were taught by 8 teachers.

The study was intended as a cluster random control trial, with the teachers from each grade randomly assigned to condition. Unfortunately, the design was compromised across grade 3 teachers. One teacher assigned to the comparison did not participate at all—project personnel did not administer pretests. Another teacher assigned to the treatment never used the intervention. There was zero uptake across her entire class. Review of email exchanges suggest three possible causes:

- The liaison between the head researcher and the school may have misrepresented the constraints of the study to the school. He reports that the school may have thought that an even number of teachers were required.

- Two gifted/high-achieving classes participated in the study. They were both inadvertently randomly assigned to the intervention. It was our intention to split these through block randomization, but we only received the pertinent data after selection and, due to a misreading of the correspondence, failed to catch the error, so no re-assignment was done. The school may have rectified our error themselves.

- It is possible that one of the teachers simply did not want to use the intervention. Project personnel doing the training reported that she attended but "had to leave early on."

Given the above, we our analyzing our study as a QED where the intact groups are the 8 classes for whom we have pretest data and the intervention group comprises those classes where any uptake occurred prior to posttest.

Teachers were provided the opportunity to indicate any students who were not prepared for fact fluency instruction. Four third-grade students were identified, 3 from the intervention group and 1 from the comparison group. These students' data were not considered as part of the study.

One of the teachers taught two classes, one within the intervention group and another in the comparison group. All other teachers taught a single class.

Group Descriptions

Table 2 provides a description of the demographic character of the groups as well as their pretest scores results. The fluency score on the pretest combines both speed and accuracy as described in the *Fluency Score Calculation* subsection.

2.2 Outcome Measures

2.2.1 Outcomes

One outcome were measured in the study: math fact fluency, which is both a key component of general math achievement and has been shown to be predictive of students' performance on general math achievement tests (see *Validity* subsection below). Fluency was measured using timed probes.

Table 2: Baseline Demographic Information

	Full Sample	Comparison Group	Intervention Group
Sample Size	129	64	65
Grade 3 Students	48.1	48.4	47.7
% Hispanic	54.2	53.1	55.3
% Asian	17.1	18.8	15.4
% White	22.5	21.9	23.1
% Black	4.7	6.2	3.1
% Multiracial	1.6	0.0	3
% Low English Proficiency	20.2	25.0	15.4
% Exceptional Student (Gifted)	25.6	25	26.1
% Free/Reduced Lunch	20.2	21.9	18.4
% Male	46.5	53.1	40
age-at-pretest (years)	8.43	8.44	8.41
pre-test % Accuracy	92.3	93.2	91.4
pre-test Speed	4.29	4.26	4.31
pre-test Score	4.58	4.57	4.58

- Grade 2 students were testing on facts with terms, minuend, and subtrahends from 0 to 10 inclusive (i.e., from $0 + 0$ up to $10 + 10$ and their associated subtraction facts.)

- Grade 3 students were tested on facts with factors, divisors, and quotients from 0 to 10 inclusive (i.e., from 0×0 to 10×10 and their associated division facts.)

These match the requirements in the *Common Core State Standards* except that, owing to that document's idiosyncratic definition of "within X" (as in "addition within 20"), a literal reading of the work indicates that facts such as $20 - 17$ and $91 \div 13$ are considered within grade level. The *Florida Math Standards* do not provide a glossary, so it is unclear whether such facts would be in the scope of the wording of its standards.

2.2.2 Probes

Probes had a format similar to those in other Curriculum Based Measurement (CBM) studies (Hintze, Christ & Keller 2002, Burns, VanDerHeyden & Jiban 2006, Stevens & Leigh 2012) as described below.

Each probe was a single-sheet of paper with 10 rows of vertically oriented problems. Probes given to grade 2 students contained addition and subtraction facts. Probes given to grade 3 students contained multiplication and division facts. The problems were printed in extra large type, so only 7 facts fit on each row. The first two rows only contained 6 facts to make room for a geometric shape placed in the upper-righthand corner to help students and monitors quickly identify which page the students were on. The problems were computer-generated with the constraint that the problems in a given row be as balanced as possible between the two operations. The facts were chosen randomly from the appropriate fact pool with each having an identical selection likelihood.

An example is provided in the *Appendix*.

2.2.3 Administrations

Three administrations were given. A pretest administration was conducted on February 12th, 2016. An interim administration was conducted on April 14th, timed to occur before heavy preparation for end-of-year testing began. A final administration was conducted on May 24th. Students were told to answer the items in order and not to skip items. The administrator used a script and was witnessed by the classroom teacher, who used a checklist to confirm each of several key points of instruction. This form also provided space for indicating any unusual occurrences.

The first and second administrations each comprised 4 one-minute fact fluency probes. Students were instructed that the first probe was a warm-up in each case. The final administration did not have a warm-up probe. It contained 3 math fact probes.

Grade 2 students also took a multi-digit computation probe, but the results of that probe were not analyzed as part of this combined report, for third grade students did not take a multidigit probe. Multidigit multiplication/division is not a core topic for third grade students in Florida and the distribution of scores on the multi-digit addition/subtraction probe were known to be fundamentally different from the distribution of scores on math fact probes, so there is no clear way to combine the two.

All students in a given grade took the same probes using the same administrative script regardless of condition. The probes that were described as “warmup” tests were not counted in any analysis.

Five students—all in comparison classes—were noted by test administrators as working on their quizzes significantly beyond the called time limit. These students were not formally considered part of the study. Posttests were taken by these students. Three of the five students scored higher on their posttest than on their pretest.

2.2.4 Fluency Score Calculation

For each student raw fluency scores were calculated as the average number of digits correct per min (dc/min) minus the number of digits incorrect per min (di/min), as this was the method found by Stevens & Leigh (2012) to have the greatest criterion validity.

Previous CBM researchers have combined grade 2 and grade 3 students (Burns et al. 2006), but to justify the pooling of their outcomes in a single analysis we conducted an analysis of the distribution of raw pretest scores for each grade separately to show similarity of distribution.

Table 3: Raw Fluency Pretest Score Distributions by Grade

Measure	Grade 2	Grade 3
Mean	20.26	20.27
Standard Deviation	10.53	11.35
Median	19	19
Kurtosis	2.37	1.91
Skewness	1.17	1.01
Range	54.67	58.33
Optimal Box-Cox (anchored at 1) λ	0.50	0.56

A Kolmogorov-Smirnov corroborated the premise that these two distributions were quite similar. It failed to reject homogeneity (critical D-stat was 0.233, calculated D-stat was 0.063, p-value = 0.99).

The distribution of these raw scores were significantly skewed and leptokurtic, as has been reported in similar studies (Burns et al. 2006), so we normalized them using a Box-Cox transformation to arrive at a final fluency score. Following the recommendation of Osborne (2005), we anchored the full distribution at a minimum value of 1 by adding 2 to all raw fluency scores. A search for an optimum λ returned 0.525, so we chose $\lambda = 0.5$ for simplicity of inversion. Thus, the calculation for final score is $\sqrt{(C - I + 2)}$, where C is the average digits correct per min and I is the average digits incorrect per min. The resulting distribution of pretest scores was not significantly skewed (skew = 0.08, SES 0.21) but was still slightly leptokurtic (Kurtosis = 0.85, SEK = 0.42). DAgostino-Pearson (p-value = 0.13) and Jarque-Barre tests (p-value = 0.13) failed to reject normality.

2.3 Validity

The criterion validity for CBM based measures in elementary math has been established by Stevens & Leigh (2012) and VanDerHeyden & Burns (2008). These studies showed math fact fluency was predictive of general math achievement on the *Oklahoma Core Curriculum* test and *Stanford Achievement Test* respectively.

2.4 Reliability

Several researchers have confirmed the reliability of CBM for math fluency.

Table 4: Previous Research on CBM Reliability for Math Fluency

Metric	Scoring Method	Source	Value
Inter-scorer Agreement	Correct Digits per Minute	(Burns et al. 2006)	0.96+
Inter-scorer Agreement	Correct Digits per Minute	(Hintze et al. 2002)	0.955
Inter-scorer Agreement	Correct Digits per Minute minus Incorrect Digits per Minute	(Stevens & Leigh 2012)	0.99+
Delayed Alternate-form Reliability	Correct Digits per Minute	(Burns et al. 2006)	0.84
Absolute Generalizability	Correct Digits per Minute	(Hintze et al. 2002)	0.75
Relative Generalizability	Correct Digits per Minute	(Hintze et al. 2002)	0.95
Test-Retest Alternate Form Reliability	Correct Digits per minute minus Incorrect Digits per Minute	(Stevens & Leigh 2012)	0.87

Our study gave 3 separate fact probes on the same day, allowing us to measure internal consistency of raw fluency score (correct digits minus incorrect digits) using Cronbach's α. The α values across the six test are described in Table 5.

Table 5: Internal Consistency of Raw Fluency Score

	Addition/Subtraction	Multiplication/Division
Pretest	0.95	0.94
Interim Test	0.96	0.94
Posttest	0.97	0.95

We also calculated delayed alternate-form reliability of the final fluency score across each grade × condition cohort and found an average value of 0.71.

Table 6: Delayed Alternate-Form Reliability (14 weeks)

	Addition/Subtraction	Multiplication/Division
Intervention	0.77	0.47
Comparison	0.72	0.89

The relatively poor value for the 3rd grade intervention group may be due to large variation in dosage. The standard variation in weekly usage across 3rd grade intervention groups was 1.24 days/week, nearly twice that of the 2nd grade intervention group, where the standard deviation was 0.65 days/week.

When dosage was added to the model predicting posttest score from pretest score, the agreement between the two intervention groups improved considerably. The coefficients of multiple correlation were $R = 0.81$ and $R = 0.77$ for the 2nd and 3rd grade intervention groups respectively.

2.5 Analytic Approach

Since randomized assignment occurred at the class level, we used an HLM modeling approach to account for cluster effects when analyzing the relationship between condition and posttest fluency. The model has two levels—grade and condition are level-2 variables, and all other covariates are level-1 variables. We used grand-mean-centered values for the lower level variables and a maximum-likelihood method for determining the random effects. If the search for a model did not converge using maximum likelihood, restricted maximum likelihood was used instead.

Models were constructed using R's *lmer* function, part of the lme4 library using the methodology for two-tier HLM models documented in a technical report from the Department of Statistics and Data Sciences, The University of Texas at Austin (UTA 2015), which showed the similarity in results to those given by SPSS, SAS, Mplus, and HLM.

We formed 3 models of decreasing complexity and calculated an effect size and statistical significance based on each.

The first model uses the same structure as that used in the original version of this report. In this model, all dichotomous and numeric covariates were used (i.e., all covariates other than race, which was polynominal), including the pretest accuracy and pretest speed. This model is most inclusive and allows for continuity between the original version of this report and the current version. It is denoted as the *Full Model*.

For the data available at the time of the original report, the pretest speed and pretest accuracy were both highly significant ($p < 0.001$). But after removing students who did not respect the time limits on the pretest or were designated as being below grade level before the study began, these additional pretest features were no longer statistically significant. A nested model χ-squared test comparing change in deviance to change in degrees of freedom did not show a statistically significant improvement upon adding either of these terms. Thus, we generated a new model lacking these two pretest features but retaining all the demographic covariates of the original. This model is denoted in the sequel as the *Demographic Model*.

In an effort to simplify the model further, we assessed the relevance of each of the demographic variables by generating a HLM with the following characteristics:

- No Level-2 variables
- Two Level-1 variables: the covariate in question and pretest score
- Group-mean-centered values
- Data scaled to be univariate

This method was selected for determining the relevance of a given level-1 factor based on Woltman, Feldstain, MacKay & Rocchi's (2012) presentation. The results are shown in Table 7. Note that this was the only analysis using group-mean centered data. The model's used for determining intervention effect and statistical significance used grand-mean centered level-1 variables.

The results of this analysis are shown in Table 7. Given their very low coefficients and t-scores, we removed gender and ESE. Upon forming the full HLM using the remaining covariates, it was found that LEP had very little impact (coefficient = 0.03) and low significance ($t = 0.24$), so it was dropped as well. In the resulting model all covariates had t-scores greater than 0.9

Table 7: Impact and significance of demographic covariates

Covariate	Coefficient	t-score
age	0.028	0.506
gender	-0.005	-0.096
LEP	0.042	0.612
Lunch	0.080	1.249
ESE	-0.006	-0.100

in magnitude and standardized coefficients greater than 0.1. There was a nearly statistically significant interaction ($t = 1.94$) between condition and whether the student was on free or reduced lunch.

This final model is denoted as the *Reduced Model*

All three models are provided in the Appendix.

Effect sizes were calculated from the coefficient for the intervention effect from each HLM-model and the pooled-within-group standard deviation of unadjusted post-test scores.

Statistical significance was determined based on the t-score of the multilevel model.

2.6 Statistical Adjustments

We used all demographic information provided except race, which was nonbinary and correlated significantly with other demographic information (R between 0.36 and 0.46 for the three most prevalent races in our sample).

Grade was coded as *grade3*, a variable equal to 1 if the student was in grade 3 and 0 otherwise.

Age was measured in years as of the pretest administration.

Gender was coded as a variable *male* equal to 1 if the student was male and 0 if the student was female.

Low-English proficiency was coded as a variable *LEP* equal to 1 if school indicated the student had low English proficiency.

Exceptional Student Status was determined based on the school's designation of the student as being within an Exceptional Student Education program. It was coded as a variable *ESE* equal to 1 if the school specified the student as belonging to an ESE program. The state of Florida specifies several ESE programs, one of which is a program for gifted students. For our study it appears this program furnished the large majority of ESE designations, as 29 of the 36 students designated as ESE were concentrated in two high-achieving classes. In grade 2 every ESE-designated student was in a single class.

Eligibility for free or reduced lunch was coded as a variable *lunch* equal to 1 if the student was eligible.

As described in the *Fluency Score Calculation* subsection, fluency was evaluated based on research-supported combination of speed and accuracy, normalized to reduce skewness via a Box-Cox transformation. This means that a student's fluency score depends on personal characteristics such as confidence, sense of urgency on a pen-and-paper test, and attention to accuracy, so students differ markedly in potential for improvement.

Pretest accuracy is the ratio of correct digits to the sum of correct and incorrect digits.

Pretest score is defined as $\sqrt{C - I + 2}$, where C is digits correct per minute and I is digits incorrect per minute.

Pretest speed is defined in a manner analogous to pretest score: $\sqrt{C - 2}$, where C is digits correct per minute. In this expression 2 is subtracted rather than added so that the expression is anchored at 1, conforming to best practices (Osborne 2005).

All student-level covariates were scaled to be univariate and grand-mean centered for improved interpretability and model convergence.

Speed, score, and accuracy on the interim administration were considered during the regression process used to impute missing data, as described in the *Missing Data* section. These metrics are calculated exactly as for the pretest using the same formula (i.e., the data was not re-anchored for the Box-Cox transformation).

An HLM model was used to calculate statistical significance for the entire sample, so no adjustment for cluster effects were necessary. We only analyzed one outcome for this study, so no adjustment was made for multiple outcomes.

2.7 Students Removed from Study

Ten students, 4 from the intervention condition and 6 from the comparison condition, were excluded from the analysis. In all cases the decision to exclude was based on information attained from the day of the pretest.

Four of these ten (3 from intervention, 1 from comparison) were excluded because their teacher indicated they were sufficiently below grade level that

Table 8: Descriptive Statistics of Control Variables

Control Variable	Mean	SD	Skew	Kurtosis
Grade3	0.48	0.50	0.08	-2.03
Age	8.42	0.57	0.03	-0.95
Male	0.47	0.50	0.14	-2.01
LEP	0.20	0.40	1.51	0.27
ESE	0.26	0.44	1.13	-0.73
Lunch	0.20	0.40	1.51	0.27
Pretest Accuracy	0.92	0.09	-2.63	8.23
Pretest Speed	4.29	1.18	0.18	0.98
Pretest Score	4.58	1.15	0.08	0.85
Interim Accuracy	0.94	0.06	-1.84	3.59
Interim Speed	5.07	1.20	0.55	0.42
Interim Fluency Score	5.31	1.18	0.46	0.39

they would not receive typical instruction in math fact fluency. This determination was provided on the day of the pretest.

Five of these ten (all from comparison) were excluded because they did not stop when time was called on the pretest. In three cases these students had higher values on their pretest than on their posttest.

One grade 2 student from the Intervention condition was noted as appearing frustrated and not working on the pretest. He had the fourth lowest fluency score of all 2nd-grade participants on the pretest and showed dramatic improvement by the interim assessment, on which he scored at the 33rd percentile within his grade. According to Reflex' internal initial testing, the student had pre-existing automaticity for 17.1% of the addition facts within 20 and had basic recall ability with 59.9%. This suggests his pretest score under-estimated his actual ability, and he was removed from the analysis for fear of artificially inflating the impact of the intervention. Note that this student was absent from the final administration.

2.8 Missing Data

Eight students, 5 from the treatment group and 3 from the comparison group, were absent for the final administration. Seven of the students had taken the interim assessment. No values were imputed for the student who missed both the interim and the final assessment. For the seven who had attended the interim test, we imputed posttest values using a multilinear regression based on students in the same instructional level group using the threshold established by Burns et al. (2006).

Table 9: Categorization of Students

Fluency (dc/min)	Category	N
Less than 14	Frustration Level	29 (22%)
14-31	Instructional Level	81 (63%)
Greater than 31	Mastery Level	19 (15%)

All available data (demographic data, pretest data, and interim test data) were used to impute posttest scores using a OLS regression that retained only statistically significant regressors.

2.8.1 Frustration Level

One of the seven students for whom posttest scores were imputed was in the frustration level. For that group, age ($t = 2.6$), pretest accuracy ($t = -3.2$), and interim fluency score ($t = 6.2$) were the statistically significant regressors.

2.8.2 Instructional Level

Six of the seven students for whom posttest scores were imputed were in the instructional level. Among students in that level, grade ($t = 4.2$), interim accuracy ($t = -2.6$), and interim fluency score ($t = 9.5$) were statistically significant.

2.9 Mastery Level

There were no students in the mastery level for whom imputation was necessary.

3 Study Data

Tables compose the large majority of this section. They are organized by table title and subsection title rather than by use of numbers.

The tables in this section report unscaled, uncentered values for ease of interpretability.

3.1 Pre-Intervention Data—All Pretest Takers

This section provides data on all students who took the pretest, including those that were formally removed from the analysis.

Outcome Data

Measure	Comparison Group				Intervention Group			
	Sample Sizes		Sample Characteristics		Sample Sizes		Sample Characteristics	
	Unit of Assignment	Unit of Analysis	Mean	Standard Deviation	Unit of Assignment	Unit of Analysis	Mean	Standard Deviation
Fluency Score	4	70	4.44	1.17	4	70	4.49	1.26

Background Data

Variable	Comparison		Intervention	
	Mean	SD	Mean	SD
Age	8.440	0.598	8.442	0.602
ESE	0.271	0.448	0.243	0.432
Male	0.500	0.504	0.400	0.493
Grade3	0.486	0.503	0.486	0.503
LEP	0.243	0.432	0.171	0.380
Lunch	0.229	0.423	0.186	0.392
Pretest accuracy	0.919	0.098	0.909	0.113
Pretest speed	4.145	1.186	4.220	1.257

3.2 Pre-Intervention Data—Baseline Sample

This section includes all students who were formally part of the analysis, including those who were absent for the posttest.

Outcome Data

Measure	Comparison Group				Intervention Group			
	Sample Sizes		Sample Characteristics		Sample Sizes		Sample Characteristics	
	Unit of Assignment	Unit of Analysis	Mean	Standard Deviation	Unit of Assignment	Unit of Analysis	Mean	Standard Deviation
Fluency Score	4	64	4.57	1.12	4	66	4.60	1.20

Background Data

Variable	Comparison		Intervention	
	Mean	SD	Mean	SD
Age	8.444	0.556	8.405	0.581
ESE	0.250	0.436	0.258	0.441
Grade3	0.484	0.504	0.470	0.503
LEP	0.250	0.436	0.152	0.361
Lunch	0.219	0.417	0.182	0.389
Male	0.531	0.503	0.394	0.492
Pretest accuracy	0.932	0.087	0.916	0.102
Pretest speed	4.268	1.147	4.324	1.213

3.3 Pre-intervention Data: Analytic Sample

Outcome Data—Analytic Sample

Measure	Comparison Group				Intervention Group			
	Sample Sizes		Sample Characteristics		Sample Sizes		Sample Characteristics	
	Unit of Assignment	Unit of Analysis	Mean	Standard Deviation	Unit of Assignment	Unit of Analysis	Mean	Standard Deviation
Fact Fluency	4	64	4.573	1.121	4	65	4.580	1.195

Background Data—Analytic Sample

Variable	Comparison		Intervention	
	Mean	SD	Mean	SD
Age	8.444	0.556	8.414	0.580
ESE	0.250	0.436	0.262	0.443
Grade3	0.484	0.504	0.477	0.503
LEP	0.250	0.436	0.154	0.364
Lunch	0.219	0.417	0.185	0.391
Male	0.531	0.503	0.400	0.494
Pretest accuracy	0.932	0.087	0.914	0.102
Pretest speed	4.268	1.147	4.307	1.214

Outcome Data—Analytic Sample with No Imputation

Measure	Comparison Group				Intervention Group			
	Sample Sizes		Sample Characteristics		Sample Sizes		Sample Characteristics	
	Unit of Assignment	Unit of Analysis	Mean	Standard Deviation	Unit of Assignment	Unit of Analysis	Mean	Standard Deviation
Fact Fluency	4	61	4.557	1.143	4	61	4.640	1.142

Background Data—Analytic Sample with No Imputation

Variable	Comparison		Intervention	
	Mean	SD	Mean	SD
Age	8.436	0.556	8.394	0.568
ESE	0.246	0.434	0.279	0.452
Grade3	0.492	0.504	0.459	0.502
LEP	0.262	0.444	0.148	0.358
Lunch	0.230	0.424	0.180	0.388
Male	0.541	0.502	0.410	0.496
Pretest accuracy	0.930	0.088	0.922	0.084
Pretest speed	4.254	1.170	4.360	1.177

3.4 Post-intervention Data and Findings

3.4.1 Analytic Sample

As grand-centered means were used for all Level-1 covariates and grade was the only Level-2 covariate other than condition, adjusted means for each group were estimated from the Constant term of the HLM model, the grade3 coefficient of the HLM model, the average value of the grade3 variable across all students, and (in the case of the intervention group) the treatment coefficient of the HLM model. Standard Deviations are unadjusted.

Outcome Data and Statistical Significance—Analytic Sample

Model	Comparison Group			Intervention Group			Estimated Effect	
	Students	adj. Mean	unadj. Standard Deviation	Students	adj. Mean	unadj. Standard Deviation	adj. Mean Difference	adj. t-score
Full Model	64	5.04	1.099	65	5.97	1.093	0.927***	5.753
Demographic Model	64	5.13	1.099	65	5.97	1.093	0.836***	4.966
Reduced Model	64	5.08	1.099	65	5.95	1.093	0.867***	4.343

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.001$

Effect size was calculated based on adjusted means, unadjusted pooled within-group standard deviations, and a correction $\omega = 1 - \frac{3}{4N-9}$ for small effect size.

Estimation of Effect Size—Analytic Sample

Model	N	*Adjusted Mean Difference*	*(unadj.) Pooled Within-Group SD*	*Effect Size (adj. Hedges' g)*
Full Model	129	0.927^{***}	1.096	0.84
Demographic Model	129	0.836^{***}	1.096	0.76
Reduced Model	129	0.867^{***}	1.096	0.79

Note: *p<0.1; **p<0.05; ***p<0.001

3.4.2 Analytic Sample with No Imputation

Analysis of students who were present for the interim assessment but absent for post assessment indicated that a full case study would substantially understate the effect of the intervention. The covariate-adjusted effect of the treatment on *interim* test scores was greater among students who missed the post test than among those who were present for all three tests. This is born out in the results of an analysis limited to those students where no imputation occurred.

Values for adjusted means for this subgroup were calculated by recentering all Level-1 covariates and generating a new HLM with the same structure as for the full analytic sample but using only those participants with no missing data.

Outcome Data and Statistical Significance—Analytic Sample with No Imputation

Model	Comparison Group			Intervention Group			Estimated Effect	
	Students	adj. Mean	unadj. Standard Deviation	Students	adj. Mean	unadj. Standard Deviation	adj. Mean Difference	adj. t-score
Full Model	61	5.07	1.12	61	5.94	1.11	0.867^{***}	5.255
Demographic Model	61	5.15	1.12	61	5.94	1.11	0.787^{***}	4.669
Reduced Model	61	5.09	1.12	61	5.92	1.11	0.828^{***}	4.249

Note: *p<0.1; **p<0.05; ***p<0.001

Estimation of Effect Size—Analytic Sample with No Imputation

Model	N	*Adjusted Mean Difference*	*(unadj.) Pooled Within-Group SD*	*Effect Size (adj. Hedges' g)*
Full Model	122	0.867^{***}	1.113	0.77
Demographic Model	122	0.787^{***}	1.113	0.70
Reduced Model	122	0.828^{***}	1.113	0.74

Note: *p<0.1; **p<0.05; ***p<0.001

3.5 Subpopulation Analyses

We analyzed sub-populations by grade. We also analyzed the sub-population of students not designated as exceptional students. Due to the smaller sample sizes, the Reduced Model was used for the analyses except grade was removed as a variable for subpopulations of constant grade.

Statistical Significance and Estimation of Effect Size

Group	N	*Adjusted Mean Difference*	*(unadj.) Pooled Within-Group SD*	*Effect Size (adj Hedges' g)*	*Adjusted t-score*
Grade 2	68	0.739**	0.94	0.78	2.46
Grade 3	63	0.877**	1.05	0.82	2.47
Non-Exceptional Students	102	0.904***	1.101	0.89	4.63

Note: *p<0.1; **p<0.05; ***p<0.001

4 Acknowledgment

We used **R**, (Team 2013) for some of the analysis in this report, especially the *lme4* package for fitting mixed models (Bates, Mächler, Bolker & Walker 2015). Other libraries utilized were *dplyr*, *tidyr*, and *magrittr* (Wickham & Francois 2016, Bache & Wickham 2014, Wickham 2016).

This document was typeset using LaTeX and makes use of the *harvard*, *booktabs*, *multirow*, *graphicx*, and *url* packages.

The *stargazer* package was used to generate LaTeX for several of the tables (Hlavac 2013).

References

Bache, S. M. & Wickham, H. (2014), *magrittr: A Forward-Pipe Operator for R.* https://CRAN.R-project.org/package=magrittr.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015), 'Fitting linear mixed-effects models using lme4', *Journal of Statistical Software* **67**(1), 1–48.

Burns, M. K., VanDerHeyden, A. M. & Jiban, C. L. (2006), 'Assessing the instructional level for mathematics: A comparison of methods', *School Psychology Review* **35**(3), 401.

Hintze, J. M., Christ, T. J. & Keller, L. A. (2002), 'The generalizability of cbm survey-level mathematics assessments: Just how many samples do we need?', *School Psychology Review* **31**(4), 514.

Hlavac, M. (2013), 'stargazer: Latex code and ascii text for well-formatted regression and summary statistics tables', *http://CRAN.R-project.org/package=stargazer* .

Osborne, J. (2005), 'Notes on the use of data transformations', *Practical Assessment, Research and Evaluation* **9**(1), 42–50.

Stevens, O. & Leigh, E. (2012), 'Mathematics curriculum based measurement to predict state test performance: A comparison of measures and methods.', *ProQuest LLC* .

Team, R. C. (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

UTA (2015), Multilevel modeling tutorial using sas, stata, hlm, r, spss, and mplus, Technical report, The Department of Statistics and Data Sciences, The University of Texas at Austin. http://stat.utexas.edu/images/SSC/documents/SoftwareTutorials/MultilevelModeling.pdf.

VanDerHeyden, A. M. & Burns, M. K. (2008), 'Examination of the utility of various measures of mathematics proficiency', *Assessment for Effective Intervention* .

Wickham, H. (2016), *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions.* https://CRAN.R-project.org/package=tidyr.

Wickham, H. & Francois, R. (2016), *dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

Woltman, H., Feldstain, A., MacKay, J. C. & Rocchi, M. (2012), 'An introduction to hierarchical linear modeling', *Tutorials in Quantitative Methods for Psychology* **8**(1), 52–69.

Appendix A Full Model

The table below describes the fixed-effects data of the full HLM for the analytic sample. This model uses grand-mean-centered values for all level-1 variables scaled to be univariate. These variables are prefixed with "c." to indicate this.

Factor	*Coefficient*	*t-score*
Intercept	4.709^{***}	25.719
c.age	0.136	0.887
c.LEP	−0.015	−0.111
c.Lunch	−0.122	−0.981
c.pre.score	−2.220	−0.932
c.gender	−0.148	−1.518
c.ESE	0.054	0.272
c.pre.speed	2.471	1.177
c.pre.accuracy	0.750	1.313
treatment	0.927^{***}	5.753
grade3	0.695^{***}	3.335
c.age.y:treatment	0.076	0.469
c.age.y:grade3	−0.224	−1.134
c.LEP:treatment	−0.021	−0.123
c.LEP:grade3	−0.027	−0.156
c.Lunch:treatment	0.195	1.440
c.Lunch:grade3	0.174	1.269
c.pre.score:treatment	1.881	1.056
c.pre.score:grade3	2.097	0.885
c.gender:treatment	0.096	0.836
c.gender:grade3	0.180	1.564
c.ESE:treatment	−0.036	−0.204
c.ESE:grade3	−0.013	−0.073
c.pre.speed:treatment	−1.327	−0.849
c.pre.speed:grade3	−1.581	−0.753
c.pre.accuracy:treatment	-0.712^{*}	−1.657
c.pre.accuracy:grade3	−0.608	−1.101
Note:	*p<0.1; **p<0.05; ***p<0.001	

Appendix B Demographic Model

The table below describes the fixed-effects data of the full HLM for the analytic sample, but excludes the pretest features of speed and accuracy. This model uses grand-mean-centered values for all level-1 variables scaled to be univariate. These variables are prefixed with "c." to indicate this.

Factor	*Coefficient*	*t-score*
Intercept	4.906^{***}	27.157
c.age	0.230	1.554
c.LEP	0.042	0.315
c.Lunch	−0.128	−1.225
c.pre.score	0.756^{***}	5.872
c.gender	−0.124	−1.283
c.ESE	0.172	0.855
treatment	0.836^{***}	4.966
grade3	0.480^{**}	2.235
c.age.y:treatment	−0.003	−0.022
c.age.y:grade3	−0.270	−1.378
c.LEP:treatment	−0.059	−0.346
c.LEP:grade3	−0.071	−0.408
c.Lunch:treatment	0.313^{**}	2.631
c.Lunch:grade3	0.174	1.459
c.pre.score:treatment	0.003	0.020
c.pre.score:grade3	0.094	0.685
c.gender:treatment	0.122	1.062
c.gender:grade3	0.120	1.039
c.ESE:treatment	−0.131	−0.734
c.ESE:grade3	−0.117	−0.642
Note:	*p<0.1; **p<0.05; ***p<0.001	

Appendix C Reduced Model

The table below describes the fixed-effects data of the HLM for the analytic sample, retaining only age,. This model uses grand-mean-centered values for all level-1 variables scaled to be univariate. These variables are prefixed with "c." to indicate this.

This model used Restricted Maximum Likelihood, as there were convergence problems when using maximum likelihood.

Factor	*Coefficient*	*t-score*
Intercept	4.827***	25.683
c.age	0.227	1.146
c.Lunch	−0.111	−0.952
c.pre.score	0.758***	5.488
treatment	0.867***	4.343
grade3	0.527**	2.219
c.age.y:treatment	−0.031	−0.154
c.age.y:grade3	−0.168	−0.692
c.Lunch:treatment	0.268*	1.936
c.Lunch:grade3	0.153	1.109
c.pre.score:treatment	−0.042	−0.286
c.pre.score:grade3	0.100	0.677

Note: *p<0.1; **p<0.05; ***p<0.001

Appendix B: Sample Addition/Subtraction Probe

9 + 8	5 + 9	18 −10	9 − 6	8 + 3	13 − 5	
10 − 3	2 + 7	3 + 8	10 + 1	12 −10	9 − 1	
5 + 0	8 + 4	3 + 6	19 − 9	7 − 1	16 −10	3 + 0
3 − 1	7 + 5	15 − 5	0 + 9	4 + 3	14 − 5	7 − 5
9 + 9	20 −10	11 − 3	4 − 4	9 + 0	6 − 1	1 +10
9 +10	12 − 3	12 − 9	2 + 8	5 − 0	9 − 4	5 + 0
4 + 4	14 − 9	7 − 0	11 − 8	7 + 0	4 − 1	6 + 5
2 + 3	12 − 5	14 − 5	4 − 4	10 +10	1 + 0	7 + 2
13 − 6	10 +10	3 + 6	9 − 6	17 − 7	10 +10	3 + 6
4 + 9	10 + 2	10 +10	3 − 0	5 + 3	5 − 5	10 −10