Strategic Reader (Project Monitor) Data Analysis

Tracey E. Hall
Michael W. Connell

Center on Online Learning and Students with Disabilities &
Center for Applied Special Technology (CAST)

April 25, 2017

The contents of this manuscript, *Strategic Reader (Project Monitor) Data Analysis*, was developed under a grant from the US Department of Education, Office of Special Education Programs (OSEP) Cooperative Agreement #H327U110011 with the University of Kansas and member organizations, the Center for Applied Special Technology (CAST), and the National Association of State Directors of Special Education (NASDSE). However, the contents of this paper do not necessarily represent the policy of the US Department of Education, and you should not assume endorsement by the Federal Government.

Introduction: Background and Context

Students with and without disabilities that have reading difficulties, which may be in decoding or comprehension, and encounter obstacles when learning by reading (Lyon, 2003). Such challenges raise considerable barriers to academic performance. Teachers are left trying to adapt curricula to meet the varied needs of their students while at the same time devising creative ways to engage *all* students (Coyne, Ganley, Hall, Meo, Murray, & Gordon, 2006). Teachers are in need of innovative supports, strategies, and tools that will make it possible to meet the educational needs of all students. By researching and developing universally designed learning environments since the 1980s, CAST has sought to fundamentally alter the relationship between children and literacy by using technology to embed reading strategy instruction directly into high quality educational content for all students. This work draws on a significant research base that supports reading strategy instruction in order to develop comprehension in students with and without disabilities (National Reading Panel [NRP], 2000; Swanson, 1999).

This paper reports on a retrospective analysis of student performance, progress, and usage in the area of language arts (vocabulary and comprehension) when supported by an online platform (Strategic Reader). The goal was to compile a unified student record from existing data, connecting demographic, usage, and outcomes measures into a single data set for analysis to explore usage patterns and academic progress and performance. The focus here was on usage of the multi-media embedded glossary available to students in the Strategic Reader platform.

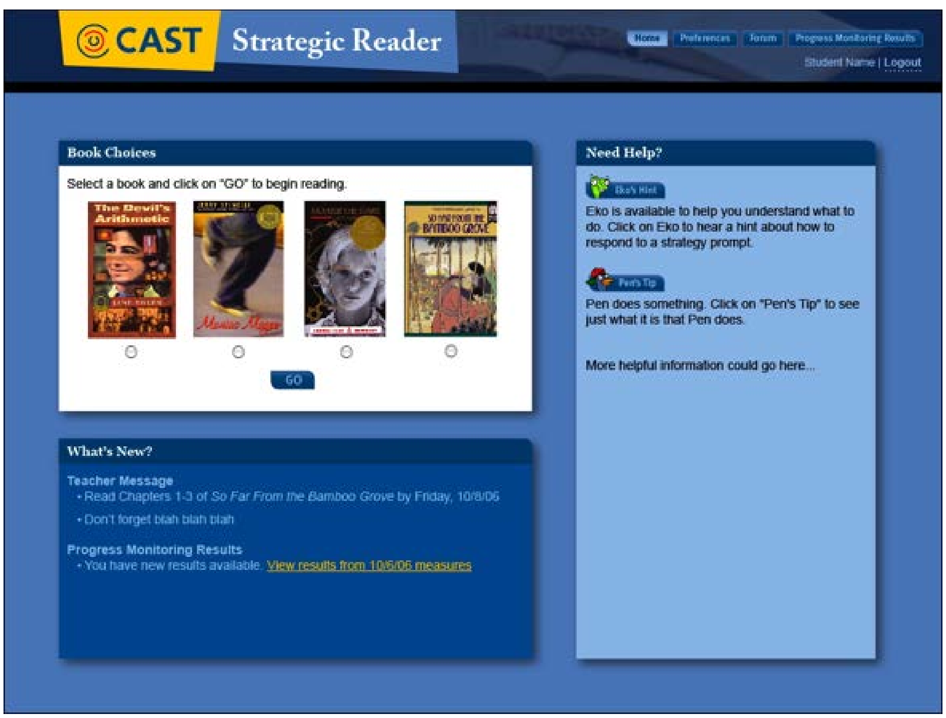
Because the study was retrospective and the component data sets were pulled from different systems, significant research was done to locate, validate, decode, and prepare the data sets for this analysis. One purpose of this report is to document the findings of this “archaeological” research for future reference. In addition, the investigation generated potentially useful insights for data management and future research.

Intervention Description

Strategic Reader is a web-based tool that was developed to research if the addition of curriculum-based formative assessment to a universally designed interactive digital reading environment would lead to better outcomes for all students, especially those students with disabilities, and modifications in reading instruction. Strategic Reader integrated formative assessment into a highly-supported literacy environment. Its reading environment allowed access for all readers to digitized texts and provided scaffolds to support comprehension. This work was predicated on the prior success of a reading environment created and researched at CAST (Dalton Pisha, Eagleton, Coyne, & Deysher, S. 2002).

Three main components were specified in the design of the CAST-developed Strategic Reader: (a) a supported and interactive digital reading environment based on UDL principles and previous research, (b) a forum for ongoing teacher-to-student and student-to-student topical discussion, and (c) embedded curriculum-based measurement (CBM) (Deno, 1985) to monitor student progress. Within the Strategic Reader tool, each component was accessible to teachers and students, with varied options and flexibility designed to meet student needs. Additionally, the teacher side of the tool contains several features and resources to support teachers' specific tool usage, such as access to aggregate and individual students' responses in each book and scores from progress monitoring, and resources to support interpretation of student data for designing instructional interventions. While the Strategic Reader tool was developed as a prototype for a prior research project, we intend to both revise and make the tool more available; at this time, there is no public access to Strategic Reader.

The Strategic Reader digital reading environment is a computer-supported environment that integrates instruction in reading strategies into high quality, age-appropriate, middle school novels in order to address English Language Arts standards on applying reading strategies to understand and interpret texts. In the digital reading environment, students had access to support features, including: text-to-speech, a dictionary, and a multimedia glossary for vocabulary in the texts available in the tool. Additionally, students respond to embedded strategy prompts as they read, and responses are recorded online in individual work logs that students and teachers can review at any time. In the reading environment, teachers could modify how much or little support a student received (with options including text highlighting and sentence starters).



Dashboard for CAST's Strategic Reader depicting navigation to books, supports, the forum notifications and progress monitoring

In the development of this project and research for Strategic Reader both quantitative and qualitative data were collected, and thus a mixed methods approach was used for our data analysis. Although randomization for this study was done by teacher, the majority of analyses were at the student level. Traditional t-tests for differences using the pre and post-test standardized measures were employed for the quantitative data. The qualitative data were coded by themes (computer experience, reading skills, forum dialogues) and then sorted into categories (e.g., navigation, supports, comfort level).

As a brief summary of the original study, students, particularly those with disabilities, who had *online* progress monitoring in the Strategic Reader (in the Treatment 1 condition) showed greater growth on reading measures than those who used the same tool without online measures. Moreover, this difference between students' performance online vs. offline was larger and statistically significant for students with disabilities, but only in the online project monitoring condition did these students show a statistically significant increase in scores (Hall, Cohen, Vue, & Ganley (2014). These outcomes sparked the additional retrospective analysis reported below to investigate if student use of the multi-media glossary were a factor in supporting student improvement.

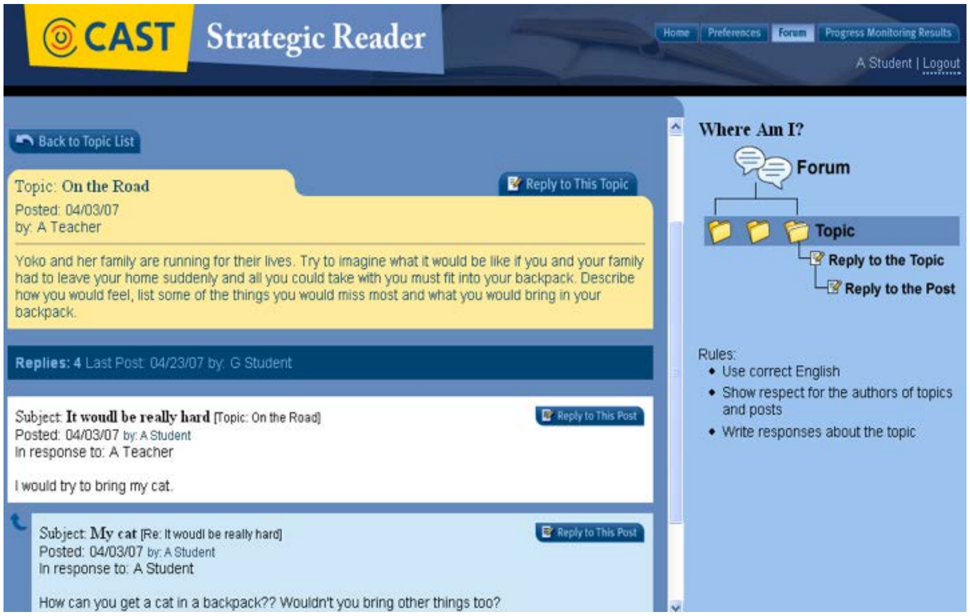
Methods

Participating students and teachers. Researchers began the original study with 14 classrooms with a total of 307 students. Parental permission and student assent was secured for 284 students. Class sizes varied greatly in each school (ranging from 17 to 36 students per class). In explanation for this variation, some classes were specialized supplemental English Language Arts (ELA) offerings for all students; others were general classroom populations of up to 36 students in integrated classes. Interestingly, these larger classes also contained mixed grade levels of students from grades 6-8. The mean age for students completing the study was 11 years six months. English was the primary language for 88% of the students; 12% spoke English and either Spanish or Portuguese. The socio-economic status and racial composition of the participant group was fairly representative of the four districts. Boys made up 50.7% of the group at 144, and girls numbered 140 (49.2%); 66% of the students were white, 20% African American, 12% Hispanic, and 2% Asian. Forty-eight percent of all participating students received free or reduced-priced lunch.

In total, 73 (25%) of our 284 student participants were identified as students with disabilities with an Individualized Education Plan (IEP). Of these students, 64 were identified as students with learning disabilities (LD) based on the district and state guidelines, eight students were identified as other health impaired (hearing and ADHD), one was identified with a hearing impairment. Six of the students with LD were also identified as having attention deficits, and three were identified for speech and language. Additionally, eight students were on 504 plans, receiving services for attention challenges (six) and mobility supports (two).

Ten teachers participated in the study, eight were female and two were male. Seven of the 10 teachers were general education ELA certified teachers; the remaining three teachers were special education teachers. The special education teachers team-taught in the general education ELA classes. All teachers were certified at the secondary level and averaged 16.8 (SD 9.7) years of teaching experience. As noted above, all teachers expressed an interest in participating in the study to use Strategic Reader in their classroom(s). Teachers were randomly assigned to condition after agreeing to volunteer for the project. All teachers participated throughout the duration of the study. We were in a total of 14 classes, four teachers assigned to employ the treatment condition, with two participating classes totaling 14, making five teachers per condition.

In the original study, the research was conducted in ELA classrooms during the regular academic school year for an 11–12-week period. During this time, students in both conditions read a minimum of two of the four available novels with identical supports and scaffolds (the digital novels contained the exact same text as the print versions). Additionally, students responded to embedded reciprocal teaching strategy prompts as they read, and their responses were recorded online in individual work logs. Class sessions were between 40 and 55 minutes in length in both treatment conditions. Teachers generally used Strategic reader three to four days each week, some (eight teachers) occasionally assigning reading and responding in the tool as homework. Students in both conditions also participated in teacher and student-directed use of the forums for student-to-student and/or student-to-teacher dialog regarding the novels. This too occurred as an expectation of class time as well as homework. Lastly, the CBM measures of oral reading fluency and maze were administered twice per month, and the reciprocal teaching strategies measure was administered at the start and completion of each novel. Students in online CBM, Treatment 1, condition completed these measures online, while those in offline CBM, Treatment 2, were administered measures in traditional paper mode.



Strategic Reader Forum where students and teachers pose questions and hold discussions about the materials they are reading.

CAST researchers provided the training for all participating teachers and students on the use of Strategic Reader. In the classrooms, the researchers observed and coached teachers and students in the use of the tool, demonstrated additional features and functions, assisted in the analysis and interpretation of progress monitoring scores, and discussed data in relation to reading strategies.

Student Assessments

Several assessments with a focus on reading ability were administered before, during, and following the Strategic Reader study to gauge growth over the 11-to-12-week intervention. Pre- and post-tests using the Gates-MacGinitie (MacGinitie, MacGinitie, Maria, Dreyer, & Hughes, 1999) standardized reading measure were administered to all subjects in both conditions. Word and passage comprehension are components of this measure.

Progress monitoring including CBM reading measures (oral reading fluency and maze), and reciprocal teaching reading comprehension strategies were developed by the researchers based on formal guidelines and procedures and administered regularly throughout the study for student and teacher use in monitoring progress and instructional decision making.

In addition, researchers developed surveys and interviews, which were conducted with participating students and teachers.

Built into the Strategic Reader tool was an event usage log, a large database designed to log the clicks students made from logging in to user clicks on the functions and features in the tool, such as reading and generation of responses and use of supports. This database was captured, sorted, and provided descriptive data on the use of features, strategies, and supports available in the tool. Specifically, student usage data of the supports and scaffolds were collected, along with responses to reading questions, teacher use, viewing of student and/or class data, intervention design, and changing levels. The event usage log is the primary data set used for this retrospective analysis, along with demographic measures and pre/post-test reading measures.

Research Questions

Two research questions guided this retrospective analysis:

- RQ1: Relationships between demographics, glossary usage, and reading performance
  - RQ1a: What is the relationship between students' use of a visual glossary during reading instruction and their performance on standardized tests of vocabulary and comprehension?
  - RQ1b: How does the relationship between glossary usage and learning outcomes differ by key demographic variables (IEP status, SES, gender, grade)?
- RQ2: Usage patterns
  - How do students' patterns of usage of the visual glossary differ by key demographic variables (IEP/Plan504 status, SES, gender, grade)?

Data

The data for this analysis included demographic, usage, and outcome measures for each student, as summarized in Table 1 below.

Table 1: Demographic, usage, and outcome measures (after renaming)

Measure Name (before → after renaming)	Type	Description	Source (see "Data file names" below)
code → code	Unique Student ID	Unique student identifier (anonymized)	Codes_Permissions
parent permission → ParentPermission	Operational (IRB)	Indicator variable, indicating if parents have signed informed consent for their child to participate in the study	Codes_Permissions Master_SS
Age → Birthdate	Demographic	Student date of birth	Master_SS
Grade → Grade	Demographic	Student grade level	Master_SS
Gender → Gender	Demographic	Student gender	Master_SS

free lunch → SES	Demographic	Student receives free lunch	Master_SS
504 → Plan504	Demographic	Student on 504 plan	Master_SS
IEP → IEP	Demographic	Student has IEP	Master_SS
type of disability → TypeOfDisability	Demographic	Type of disability	Master_SS
Stanine Voc → PreVocStan	Outcome	Pre-test vocab score (Stanine)	Master_SS
Stanine Comp → PreCompStan	Outcome	Pre-test comprehension score (Stanine)	Master_SS
Stanine Voc.1 → PostVocStan	Outcome	Post-test vocab score (Stanine)	Master_SS
Stanine Comp.1 → PostCompStan	Outcome	Post-test comprehension score (Stanine)	Master_SS
NCE Voc → PreVocNCE	Outcome	Pre-test vocab score (NCE units)	Master_SS
NCE Comp → PreCompNCE	Outcome	Pre-test comprehension score (NCE units)	Master_SS
NCE Voc.1 → PostVocNCE	Outcome	Post-test vocab score (NCE units)	Master_SS
NCE Comp.1 → PostCompNCE	Outcome	Post-test comprehension score (NCE units)	Master_SS
inserttime	Usage	Date and time event was logged	Event_log
type	Usage	Type of event (login, glossary, etc.)	Event_log
detail	Usage	Additional information associated with the event	Event_log
person	Usage	ID number (not sure what it references)	Event_log
loginsession	Usage	<Not applicable>	Event_log
form	Usage	<Not applicable>	Event_log
wordid	Usage	<Not applicable>	Event_log
textid	Usage	Name of book	Event_log
chapter	Usage	Chapter number of word if glossary event	Event_log

page	Usage	Page number of word if glossary event	Event_log
level	Usage	<Not applicable>	Event_log
strategy	Usage	<Not applicable>	Event_log
task	Usage	<Not applicable>	Event_log
word	Usage	Word looked up (if glossary event)	Event_log
Derived variables			
DeltaVocNCE	Outcome		PostVocNCE-PreVocNCE
DeltaVocStan	Outcome		PostVocStan-PreVocStan
DeltaCompNCE	Outcome		PostCompNCE-PreCompNCE
DeltaCompStan	Outcome		PostCompStan-PreCompStan

Data file names

Codes_Permissions: 2008_2009 student codes and permission status_ALL.xls
Master_SS: Master data spreadsheet PRoj MOn.xls | sheet=Proj Mon '08_'09
Event_log: ProjMon2_FirstLastEvents.csv

Values of 'type' field in our sample

The 'type' field in the events log contains information about the type of user event being recorded (logging in, viewing the glossary, etc.). For this analysis, the events of particular interest were those related to viewing the glossary, where type==view:definition.

Data Preparation

The data preparation process involved several steps.

Demographic and Outcomes Data

The demographic and outcomes data were collected a number of years before the current project was initiated. Researchers on that project compiled, cleaned, and prepared the data for analysis, and we had access to their prepared data file (called "Master data spreadsheet PRoj MOn.xls" in the worksheet titled "Proj Mon '08_'09 "). Listed below are the major steps we took to prepare these data for the current analysis.

- Drop unused columns. Many columns in the spreadsheet with column headings (field names) had no data. These empty columns were all deleted. The "Age" column was coded inconsistently—it was called "Age," but it was actually in most cases a date, presumably the student's date of birth; in other cases it was an integer—probably the age, though it's not clear how or when that was calculated. It was done before the spreadsheet was handed off, and the spreadsheet doesn't contain a "test date" field we might use to calculate age at test time. Since we had the student's grade, and that tends to be a better predictor anyway, we dropped the Age column (which was first renamed “Birthdate” and then dropped later).

- Rename columns. Columns were renamed for a number of reasons. Some column names were long. For example: “Treatment (1=online, 0=offline)”. This is not a problem—in fact, it’s helpful to have information about how to interpret the field right in the field name. But for working with the data during analysis, it’s easier to work with a shorter name. Because of the way the spreadsheet was set up, some columns when read in for analysis had duplicate names (like the matching pairs of pre and post-test scores). These were renamed to make them distinct and to make it clear from the name what each one was.
- Drop records. Drop records without parent permission and records without demographic data.
- Fill in missing values. The fields Plan504, IEP, and SES were coded ‘y’ for ‘yes’ and had a missing value for ‘no’; we imputed ‘n’ for the missing values.
- Add derived variables. Added delta scores on the performance measures (post-test scores minus pre-test scores for vocabulary and comprehension for each of NCE and Stanine scales).
- Validate data. Check for duplicate student codes (there were none).

After cleaning, 214 student records remained in the demographic/performance data set.

Usage data

- Create merge key. The usage data set did not have the same unique student ID as the performance/outcomes data so we had to create a full name field from the first and last name to merge on. Merge usage data with unique student ID. The full name field was used to merge the usage data table with a table containing the full name, the unique student ID, and the parent permission.
- Drop rows. Rows corresponding to student records with no parent permission were dropped.
- Normalize text fields. The “word” field had “word=” preceding the actual word that the student looked up in the glossary. The “word=” text was stripped off of each entry, leaving just the target word.

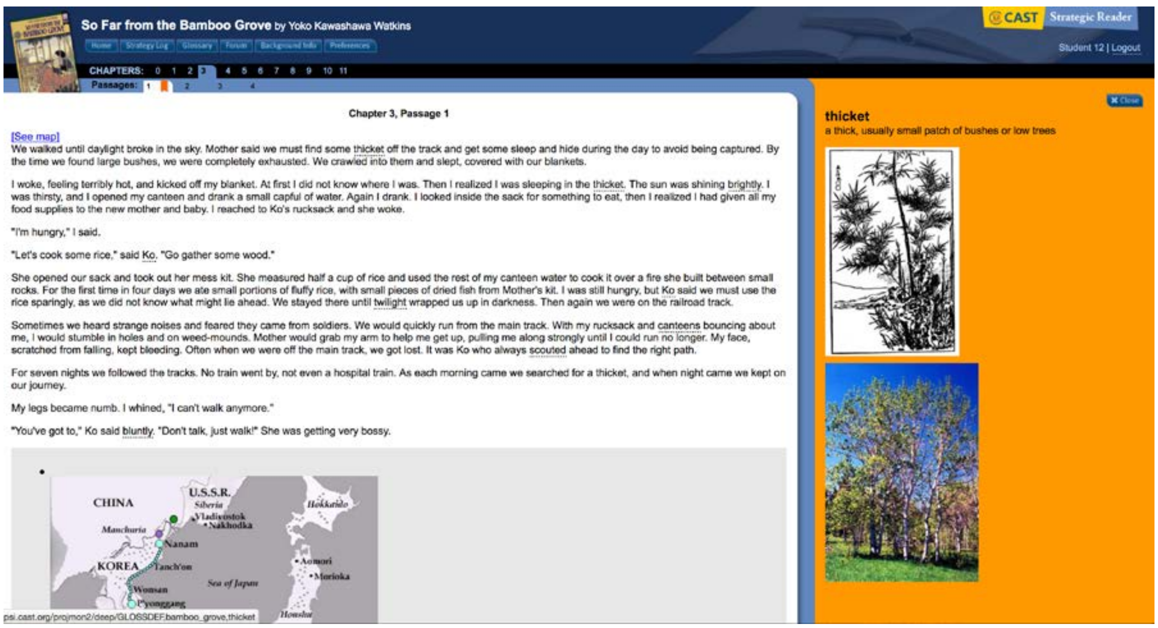
After cleaning, usage records for 220 students remained in the data set.

Unified Data Set

- Merge demographics/outcomes data set with the usage data set. Drop unused columns. Dropped parent permission field at this point, since we had already dropped the records for students without parent permission
- Normalize data. Mapped fields with ‘y’ and ‘n’ values for yes and no (Plan504, SES, and IEP) to indicator values (1 and 0).
- Select records for glossary definition usage events only. There were 114 students who did not use the glossary at all.

- Transform variables. The count of glossary uses per student was skewed right. This variable was log transformed to create a more normal distribution that could be used in regression analyses.

After cleaning, unified records for 259 students were in the data set. After selecting records for glossary definition usage events, the number of unified records were for 145 students in the data set (114 students who had no glossary usage were dropped from the analysis). After removing records for students who had no demographic or outcomes data, unified records for 117 students remained (28 students with no demographic or outcomes data were dropped from the analysis).



Vocabulary words are indicated with dotted underline. When selected, a multi-medial definition appears in the right margin.

Analyses

Exploratory Analysis: Univariate plots

Outcome variables (change scores for vocab and comprehension)

The change scores for the reading comprehension and vocab outcome variables for the two scaled measures (NCE and Stanine) are shown in Figure 1. All look reasonably normal as expected.

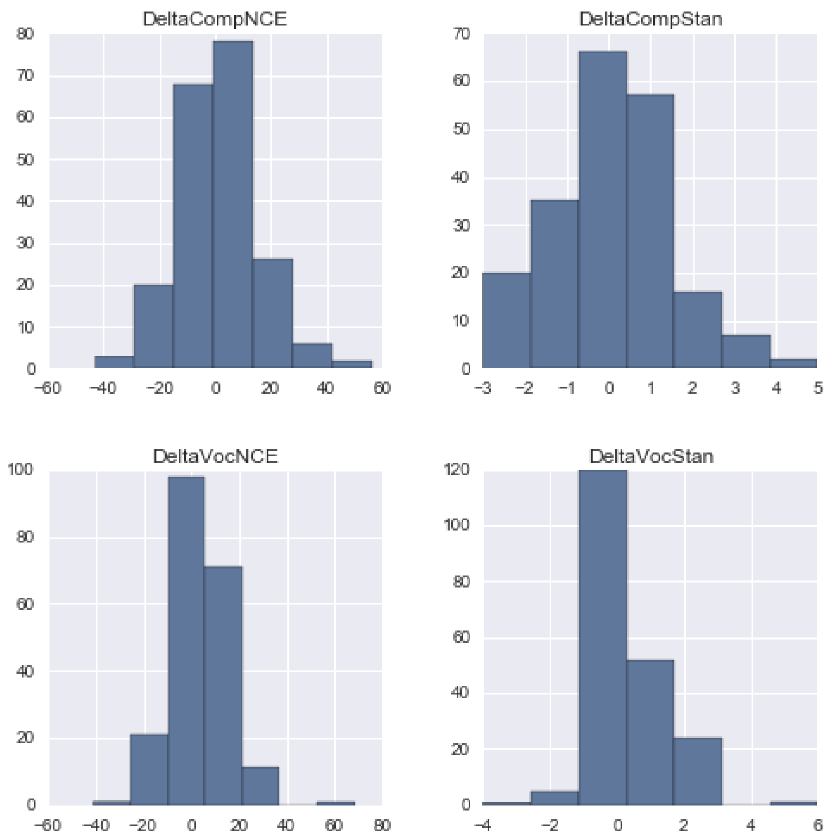


Figure 1: Univariate distributions of change scores (post-pre) for the vocabulary and comprehension assessments used to measure student performance.

In addition, paired-sample t-tests were used to look at whether the post-test scores on each measure were significantly different from the pre-test scores at the 0.05 significance level. The score changes for vocabulary tests were significant, but the score changes for comprehension were not (see Table 2).

Table 2: Results of t-tests comparing pre- and post-test scores on vocabulary and reading comprehension using two scales (NCE and stanine).

Construct	Scale	p-value
Vocabulary	**NCE**	**1.86e-06*****
Vocabulary	**Stanine**	**0.00049*****
Reading Comprehension	NCE	0.83
Reading Comprehension	Stanine	0.23

*** p < 0.001

Predictor variables

The main predictor is the count of how many glossary words a student looked up (glossary_usage_count). For this count, the 'view:definition' type events were counted, as this event indicates the number of times a student accessed a glossary definition. A histogram of the count of glossary events for each student (n=117) shows that this distribution is skewed right (Figure 2).

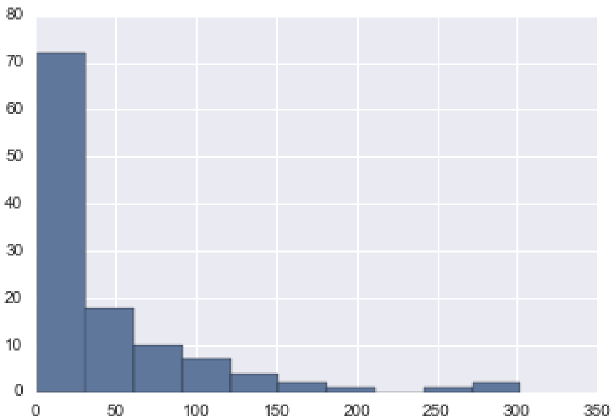


Figure 2: Histogram of the count of glossary events for each student in the sample (n=117).

For analysis, it would be better if the predictor were normally distributed. A log transformation was used. As expected, the log transformation gives a more normal looking distribution (see Figure 3, n=117), although there seems to be a long tail on the left, and it appears to be bimodal with the secondary mode near zero. This bimodality may be an artifact of the implementation. For example, if teachers required students to use the glossary a certain minimum number of times as part of the requirements, or if a tutorial demonstrated how to do it, students might use it a small number of times immediately following this demonstration and then stop. Further, if students were exploring they might use it a few times before the novelty wore off. Since the log-transformed version of the predictor is more normal than the original variable, it will be used in all analyses.

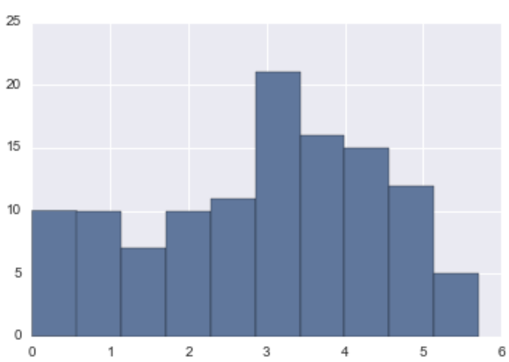


Figure 3: Histogram of the log-transformed count of glossary events for each student in the sample (n=117).

Exploratory Analysis: Bivariate plots

Outcomes vs. Usage Predictor

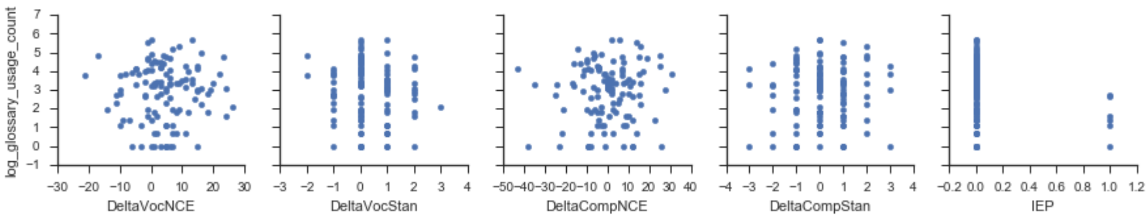


Figure 4: Bivariate plots of usage vs. each performance measure (change in vocabulary and comprehension, NCE and Stanine scales), as well as usage by IEP status (0=no IEP).

Visual inspection of bivariate plots between main predictor (log_glossary_usage_count) and each outcome variable (see Figure above) show a small number of points on the edge of the data cloud. These data points do not look extreme or particularly unusual for the DeltaVocNCE or DeltaVocStan outcome variables (other than some bunching up around zero). For the DeltaCompNCE and DeltaCompStan variables, some structure near the low values of log_glossary_usage_count is noted. Looking back at the univariate distribution of that variable, it seems to correlate with the bimodal structure of the data. The transformed variable is more normal than the original so we will use that one for all analyses going forward.

Correlations

The correlations between predictors and the outcome variables are all quite low (< 0.10) - see Table 3. Yellow highlighting is used to draw readers' attention to most important cells.

Table 3: Correlations between usage variables and performance variables.

	DeltaVocNCE	DeltaVocStan	DeltaCompNCE	DeltaCompStan	glossary_usage_count	log_glossary_usage_count
DeltaVocNCE	1.00	0.88	0.12	0.13	0.00	0.01
DeltaVocStan	0.88	1.00	0.01	0.01	-0.08	-0.09
DeltaCompNCE	0.12	0.01	1.00	0.95	0.09	0.07
DeltaCompStan	0.13	0.01	0.95	1.00	0.04	0.06
glossary_usage_count	0.00	-0.08	0.09	0.04	1.00	0.77
log_glossary_usage_count	0.01	-0.09	0.07	0.06	0.77	1.00

Linear Regression Analysis

Multivariate linear regression was used to model relationships between demographics, student usage of the visual glossary, and reading outcomes (Table 4).

Table 4: Nested hierarchy of linear regression models relating demographics, student usage of the visual glossary, and reading outcomes (n=113). Yellow highlighting is used to draw readers' attention to interesting results.

Outcome variable	Intercept	log(glossary)	PreCompNCE	PreVocNCE	IEP	IEP x PreCompNCE	R-squared
DeltaVocStan (n=113)	0.47^{*}	-0.05					0.008
DeltaVocNCE (n=113)	4.00^{*}	0.08					0.000
DeltaCompStan (n=113)	0.009	0.046					0.003
DeltaCompNCE (n=113)	-1.47	0.60					0.005
DeltaCompNCE (n=113)	11.71^{**}	1.77^{*}	-0.27^{***}				0.137
DeltaCompNCE (n=113)	4.76	1.39^	-0.61^{***}	0.47^{***}			0.320
Best model: DeltaCompNCE (n=113; Note the contributions of IEP and the interaction are small compared to the main effect of the pre-test variables and the main predictor, and there are only 6 students on IEPs in this sample)	1.74	1.52^{*}	-0.60^{***}	0.49^{***}	37.57^{*}	$-0.95^{\wedge}$ (marginal at 0.05 level)	0.347

^p<0.1, * p < 0.05, ** p < 0.01, ***p<0.001

When controlling for pre-test scores on vocabulary and reading comprehension, as well as the IEP status as a demographic variable, and including the interaction between IEP status and the pre-test reading comprehension score, student usage of the visual glossary is found to have a significant positive association with the change in reading comprehension scores when represented using the NCE scale. Specifically, exponentially greater usage of the visual glossary (going up by a factor of about 2.7), is associated with linearly greater increases in reading comprehension in increments of about 1.52 points.

While this finding may suggest promising directions for future research, it is not robust in this sample. In particular, the additional variation associated with IEP and the IEP x PreCompNCE interaction (which together contribute an incremental increase to the R-squared of 0.027) are small compared to the variation associated with the background variables (R-squared=0.320), the interaction between IEP and PreCompNCE is marginally significant, and removing that interaction makes IEP non-significant.

Sensitivity analysis

To explore whether the bimodal-looking structure in the predictor variable affected the results, we ran a series of regressions, dropping the low points where the second mode appears to be. Running the same regressions for log_glossary_usage_count against each of the four outcome variables and dropping the data points where log_glossary_usage_count < 2.0 did not change the linear regression results qualitatively.

Are there differences in usage patterns between students with and without disabilities?

The original sample had 74 students with IEPs; as a result of the data cleaning, removing students with incomplete data sets, and students with no usage of the vocabulary supports, the final population of students with IEPs in the analytic sample was six. Overall students with IEPs used the visual glossary less on average than students not on IEPs. Their usage ranged from 1-15 uses of the glossary, with counts of one, three, four, five, 14, and 15 total uses. The limited sample size narrows the analytics for this research question.

Are there differences of usage by demographics and pre-scores?

Additional regression models were run to examine whether different demographic groups of students exhibited different levels of glossary usage (Table 5). On average:

- Students on IEPs used the glossary significantly less than students without IEPs
- Students in higher grades used the glossary significantly less than students in lower grades
- Low-SES students used the glossary significantly less than other students
- Students with higher pre-test scores on Comp and Voc (NCE scale) used the glossary significantly more than students with lower pre-test scores
- No association was found between Gender or Plan504 and glossary usage

Table 5: Regression models examining relationships between each demographic variable and glossary usage. Yellow highlighting is used to draw readers' attention to most interesting results.

Outcome variable	Intercept	Predictor	R-squared
log_glossary_usage_count	2.99***	IEP: -1.42*	0.043
log_glossary_usage_count	7.41***	Grade: -0.70***	0.122
log_glossary_usage_count	3.01***	Gender: -0.21	0.005
log_glossary_usage_count	2.90***	Plan504: 0.97	0.007
log_glossary_usage_count	3.31***	SES: -1.32***	0.161
log_glossary_usage_count	1.12*	PreCompNCE: 0.0286***	0.128
log_glossary_usage_count	1.20*	PreVocNCE: 0.0271***	0.110

^p<0.1, * p < 0.05, ** p < 0.01, ***p<0.001

Discussion

In this secondary analysis of existing data collected for a prior study, we were able to compile a unified student record, including student demographic data (IEP), student usage data (visual glossary usage), and student outcomes (pre- and post-test scores on vocabulary and reading comprehension).

The best regression model coming out of our analysis suggests that students who made greater use of the glossary showed greater increases in reading comprehension on average, controlling for pre-scores on comprehension and vocabulary, IEP, and the interaction between IEP and reading comprehension pre-score. The association does not seem to be robust—in particular, the interaction term is marginally significant, and if we remove that term the main predictor becomes marginally significant. Also, the incremental variation associated with the main predictor and with the IEP and interaction terms (as measured by changes in R-squared) is small. This lack of robustness may be at least in part due to the small number of students on IEPs in this sample. This finding is nonetheless promising and points to potential design features of future studies, namely:

- Use the reading comprehension measure on the NCE scale as an outcome measure (it was the outcome variable with the best analytic properties — containing more information than the Stanine and having an interval scale unlike the percentile rank—and therefore the only one exhibiting statistically significant effects in this sample).
- Design future studies to ensure more students with IEPs are represented in the sample—the IEP coefficient is significant, and the interaction between IEP and the pre-test score on comprehension is marginally significant, even though there are only six students on IEPs in this sample. A larger sample would help establish whether this is a robust and generalizable relationship.
- The findings might be strengthened if steps were taken to ensure more uniform implementation of the intervention in classrooms (or a measurement of implementation were taken, at least). This change to the implementation would likely require defining what the ideal implementation is, and two clues coming out of the present analysis may be that: 1) authentic, repeated use of the glossary may support increased reading comprehension, and 2) exponential increases in usage lead to linear increases in reading comprehension, and so substantially larger levels of student usage of the glossary may be needed to produce a larger effect that could be detected in the analysis. Systematic training for teachers and students might improve results, as well as requirements on how many times students must use the glossary (as long as teachers ensure students are really using it and not merely going through the motions to check the box in the requirements).

Some groups in this study (students on IEPs, low SES students, students in higher grades, and students with lower pre-test scores on vocabulary and comprehension) used the glossary significantly less than others. In other words, students who might find the scaffold most helpful evidently tend to use it the least. Together with the finding of the present study that greater glossary usage is associated with improved reading comprehension outcomes, this finding suggests that future studies of this intervention should increase the emphasis on training students to use the tools, training teachers to support students in using the tools, and fidelity of implementation (perhaps with requirements for students to use the tools at least a certain number of times). The teacher training might include particular tactics for students who might be least inclined to use the vocabulary supports and scaffolds. In preparation for future studies, a literature review on student use of help systems and hints might yield some usable insights on this front.

As with many retrospective analyses—and especially analyses that synthesize multiple data sets originally collected for disparate purposes—the process of compiling and preparing the data for analysis itself required significant effort (Humphrey, 2016). Major challenges are often introduced when data are re-used for secondary analyses or new purposes. Many of the challenges arise because information is lost at each stage of the original research (Humphrey, 2006). For example, when survey data are entered into a database system, important information may be lost about the original rationale for the survey design, the scale used, and any transformations that were done on the data (for example, reversing the "polarity" of questions framed in a negative way so the responses to those questions align with the polarity of the positively framed questions). To take another common example, when a secondary researcher acquires data from the original researcher on a project, often the names of database fields cannot be interpreted immediately without some kind of dictionary explaining them or without talking to the original researcher (who may or may not remember their exact meaning). The most significant challenges encountered in this research project included:

- Ambiguous or obscure names for fields in the data set(s)

- Lack of a "data dictionary" explaining how to interpret fields, especially fields containing symbolic codes and/or complex information (Example: the "detail" field in the usage logs contained important information the interpretation of which depended on the values of other fields)

- Obscure relationships between elements and records across different data sets (for example, it wasn't clear how the unique student codes were generated in one file and, therefore, how the student records in that file could be linked to the corresponding student records in other files).

- Ambiguous or obscure codes used to specify the values of data fields—for example, for the "IEP" field some entries were 'y', some 'n,' and some missing; and it wasn't clear if the missing values should be interpreted as missing or interpreted as "no IEP" (same as 'n'). Inconsistent data coding—for example, the "Age" field in the master spreadsheet sometimes contained an integer and sometimes contained a date .

Recommendations for integrating multiple data sets

This study was not a typical research project in the sense that the data used were not collected as part of this investigation, nor was the data collection designed specifically to support our research questions. In addition, this research is not a typical secondary analysis of an existing data set. In this study, we sought to compile *multiple data sets* collected for *different purposes* into a single unified data set in order to understand how student demographics, usage, and learning outcomes relate—to conduct a post hoc investigation. As a result, the process of compiling the unified data set and researchers conducting this analysis themselves generated some insights that might prove useful for future research of this sort—in particular, in relation to a) data discipline (data collection and management), b) research documentation, and c) research design and implementation.

A) Data discipline

a. Develop a data management plan as part of the research planning phase (ICPSR, 2012) that specifies what data are to be collected; how they will be stored, documented, and archived; and who will be responsible for managing and documenting them, with the goal of making the data archive self-standing and re-usable by future researchers without additional support from the original research team

b. Specific, straightforward best practices that would have helped in this investigation and are generally useful include:

    i. Design database and data storage systems to use data validation and best practices, such as storing each piece of data only once in the system and referencing it multiple times if it is used in more than one place (e.g., student name—don't enter it into multiple tables or even databases)
    ii. If possible, issue a common student ID and use it in all systems related to research; barring that, store a unique combination of values (name plus grade or birthdate) in each location so the records can be unified after the fact, and document this in cases where privacy concerns are an issue
    iii. Record all values of a variable explicitly. For example, don't use "null" or "missing" values to carry meaning—such as "y" for yes and "missing value" for "no." There's no way to distinguish between (invalid) missing values and (valid) entries in that case.

B) Research Documentation

a. Record definitions of each table in the database and each field in each table; this information can be stored right in the database by creating a number of separate tables that associate codes with their human-readable definitions (preferable) or in a text or other file stored alongside the data source (less preferable because the files can get separated)
b. Specify the meanings of all possible values of indicator and categorical variables (e.g., Gender: Male = 0? Male = 1?; IEP: 'y' = 'yes', null = 'no'?); this can be done in the field name itself in a separate table (preferable, for data validation reasons) or in a separate file (least desirable, as the files can become separated)
c. Document how any derived variables are produced

d. Document any information people need in order to re-generate data, understand where the data came from, interpret the data, check assumptions, and conduct additional analysis on the data, even on research questions not related to the original questions
e. Make the data as self-documenting as possible. Whenever possible, store metadata as close to the data as possible and in ways that it is least likely to get separated— for example, using field names to document categorical values such as "Gender (0=Male, 1=Female)" as the field name for a Gender field makes that field self-documenting
f. As part of the research process, maintain documents that capture descriptions of the provenance of data—how they were collected, copies of the instruments used to collect them, any transformations done on the data, procedures for creating derived variables, and data cleaning procedures (what was done to missing data, for example, and why)
g. When possible, use a method that retains the original data unrevised in the original data source archived, along with the scripts and processes used to generate the analytic data set (e.g., through a Python or R script); although note that this makes the data less portable (additional tools and expertise may be required to make use of the data in these cases), and careful thought should be given to what tools (open source vs. commercial), what expertise (knowledge of specific programming languages), and other special requirements it might place on a secondary researcher

C) Implementation

A major issue related to implementation is that increasingly, data that are collected for one purpose (e.g., operating a web site) are being used for other purposes (e.g., research on system usage patterns and user behavior), and it would facilitate future research if web logs were designed to accommodate the possibility of future research whenever possible. In addition to the recommendations above regarding data discipline and documentation (especially documenting the meaning of events and other data stored in usage logs so they can be understood and interpreted by later researchers who may not have access to the original working platform), new designs for user interfaces and event logging protocols allowing one to readily recover user behavior would be ideal. For example, it might be sufficient to record a single generic event using a feature such as "button click" whenever a user presses one of a number of buttons; however, such a feature makes determining what action the user actually took after the fact difficult or impossible if additional context is not provided. On the other hand, if a single button can perform different functions depending on context (such as how the user reached the page the button is on), those different uses should probably be associated with distinct events if those events encode important usage information that needs to be disentangled later.

Some groups of students (including some of those who might benefit most from using learning scaffolds) seem to use the scaffolds least. Generally speaking, it might help if teachers could make extra effort to "normalize" usage and encourage it among all students (and make sure these students are using it at similar rates to other students). Some specific ideas for how to address this issue include:

a. Provide training for teachers on how to introduce the supports to students to make sure they are aware they exist and are motivated to try them out/use them over the long term
b. Design and provide incentives for students aligned with appropriate usage (see the literature on gamification and help-seeking behavior for examples)
c. Have teachers record "training days" when students are explicitly shown how to use the supports so researchers can take this into account in the analysis (specifically, usage will be higher on training days, and usage may not be entirely authentic if it is required to fulfill a formal requirement)
d. Have teachers set a minimum number of uses required by each student (and for purposes of data management and future analysis, record implementation details related to usage such as minimum number of uses required)
e. Represent student participation in their learning of the tool and features so that such information can be incorporated into moderator analysis.

Conclusions

Retrospective analysis gives rise to special challenges, especially when data sets collected for one purpose (e.g., operating an online platform) are used for another (e.g., modeling the relationship between demographics, usage patterns, and outcomes). As this study shows, however, promise exists in this kind of research. Such research could take at least two forms. First, while it stands to reason that the chances of producing robust findings are likely to be reduced when data are used "off label," as they were in this study, the fact that we found marginally significant results in support of new research questions suggests that this kind of research can, at the very least, point toward promising avenues for future research studies. The findings suggest what measures to use and how to implement an intervention. Second, the process of producing a unified data set plus analyses of that data set can yield a wealth of insight about how to improve measures, standardize operating procedures, data management, implementation, and other aspects of research that would enable researchers generally to extract more value from research data. For example, as web log data are likely to be used more frequently over time for research purposes, it would make sense to develop new guidelines and disciplines around the design of web logs, which would improve the usability of the data for research purposes, especially where such design choices do not interfere with the primary operational function of the web log and do not add significant overhead to the implementation process. Even some simple improvements in data management and documentation across the field would likely create many new opportunities to increase the value extracted from data collected for research purposes.

References

Coyne, P., Ganley, P., Hall, T., Meo, G., Murray, E., & Gordon, D. (2006). Applying Universal Design for Learning in the classroom. In D. Rose & A. Meyer (eds.), *A Practical Reader in Universal Design for Learning* (pp. 1-13). Cambridge, MA: Harvard Education Press.

Dalton, B., Pisha, B., Eagleton, M., Coyne, P., & Deysher, S. (2002). *Engaging the text: reciprocal teaching and questioning strategies in a scaffolded digital learning environment.* Final report to US Department of Education, Office of Special Education Programs.

Hall, T. E., Cohen, N, Vue, G. & Ganley P. (2014). Addressing learning disabilities with UDL and technology: Strategic reader. *Learning Disabilities Quarterly,* Online publication DOI:10.1177/0731948714544375

Humphrey, Charles. (2006). "e-Science and the life cycle of research." Retrieved December 13, 2016, from https://www.usit.uio.no/om/organisasjon/uav/itf/saker/forskningsdata/bakgrunn/life-cycle.pdf

Inter-university Consortium for Political and Social Research (ICPSR). (2012). *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle* (5th ed.). Ann Arbor, MI.

Lyon, G. R. (2003). Reading disabilities: Why do some children have difficulty learning to read? What can be done about it? *Perspectives, 29*(2), 17-19.

Additional resources on data management planning

Data Management Planning overview and tools:
http://libguides.bc.edu/dataplan/overview

Data Documentation Initiative:
http://www.ddialliance.org/explore-documentation