AERA Open
January-December 2021, Vol. 7, No. 1, pp. 1–20
DOI: 10.1177/23328584211025528
Article reuse guidelines: sagepub.com/journals-permissions

Improving Outcomes for English Learners Through Technology: A Randomized Controlled Trial

David Harper
Anita R. Bowles
Lauren Amer
Rosetta Stone

Nick B. Pandža
Jared A. Linck
University of Maryland

*English learners (ELs) in K–12 schools must acquire English while simultaneously mastering content knowledge. Educational technology may support students' learning through the affordance of individualized language practice. The current randomized controlled trial intervention study examined the effects of Rosetta Stone Foundations software on English learning among middle school ELs. The study took place in Grades 6 to 8 of an urban U.S. school district (*N *= 221). Predictors of interest included time of testing (pretest vs. posttest) and software usage, and covariates included grade level, sex, and attendance. Additionally, socioeconomic status and home language were accounted for due to sample homogeneity. Multilevel models indicated that treatment group students showed larger gains than control group students on oral/aural outcomes. These results indicate that the software intervention enables individualized practice that can produce proficiency-related gains over and above the typical classroom curriculum.*

Keywords: *bilingual/bicultural, classroom research, computer applications, computers and learning, ELs, English learners, experimental research, hierarchical linear modeling, intervention, language comprehension/development, middle schools, multilevel models, randomized controlled trial, RCT, Rosetta Stone, software, statistics*

The percentage of English learners (ELs) in U.S. K–12 schools is increasing, and, by 2016, approximately 4.9 million students fell into this classification (McFarland et al., 2019). For ELs, acquiring English while simultaneously mastering content knowledge can be challenging. Consequently, ELs show a consistent achievement gap when compared with students who enter school already fluent in English (Fry, 2008). Educators are often not adequately trained to address the needs of the EL population (National Academies of Sciences, Engineering, and Medicine, 2017), and there is a shortage of teachers with specialized training for this population (Sutcher et al., 2016).

Educational technology may improve this situation by providing ELs with differentiated instruction, multimodal lesson content, and other types of digital support features (U.S. Department of Education, 2018). In addition, by lowering the affective filter, that is, by reducing students' anxiety around learning, through the use of technology, some of the obstacles to speaking the language may be diminished (Krashen, 1982). Studies have shown that increased opportunities for language production practice, in particular, may lead to improvements in several learning domains and may also improve retention (Boiteau et al., 2014; Hopman & MacDonald, 2018; Roediger & Karpicke, 2006). Additionally, the increased speaking opportunities that technology provides may help to combat the problem of student reticence documented in the English as a Second Language (ESL) classroom (Donald, 2010).

To our knowledge, there are few, if any, randomized controlled trials (RCTs) investigating the effectiveness of a software intervention on all four language skills, that is, reading, writing, listening, and speaking, in EL populations. Although studies often include EL status as a variable, they generally only focus on the effectiveness of a literacy intervention (see Cheung & Slavin, 2012; Richards-Tutor et al., 2016) or include receptive skills only (e.g., Troia, 2004). Controlled studies that assess an intervention's comprehensive effects on all four language skills in a K–12 context are urgently needed.

Purpose of Study

Although there is an abundance of evidence that computer-assisted language learning can have a positive impact on

student outcomes (for reviews, see Golonka et al., 2014; Plonsky & Ziegler, 2016), the effectiveness of an intervention can depend on many factors, such as methods of implementation, student age(s), usage time, and so on. For low-proficiency and intermediate ELs in a K–12 setting, the value of individualized practice with technology has not been rigorously investigated, despite thousands of learners around the country using such interventions every day in the classroom. Perhaps two of the more appealing and distinctive features of such software interventions are that they provide language students with more opportunities to practice speaking, and students can learn at their own pace. These features may be particularly important for beginning learners who need extensive, targeted instruction in the basics of the English language prior to engaging with activities in the larger classroom. Speech practice within an e-learning environment allows novice learners to practice English language output repeatedly, receive automated computerized feedback, and build confidence without fear of judgment by peers or teachers. An additional benefit of e-learning software is that students can proceed at their own pace. And, as of this writing in March 2021, many students across the country are forced to learn from home due to coronavirus disease 2019 (COVID-19); e-learning software can be an essential practice tool for learners who are no longer in the classroom and speak a language other than English at home. The purpose of this study is to examine whether incorporating a software package affording these features translates to increased student achievement compared with a business-as-usual classroom curriculum.

This study investigated the effectiveness of Rosetta Stone Foundations software as part of a blended curriculum, broadly defined as a mix of face-to-face and online learning environments (Stacey & Gerbic, 2009). Our primary research question was "Is there a relationship between Rosetta Stone Foundations usage and standardized test scores?" Specifically, we evaluated the software's effect on student achievement as measured by the Pearson Test of English Language Learning (TELL; Bonk, 2016). The TELL is aligned to state standards on English language development (see the Materials section for more information) and has three different test types. The version of TELL used for this study was the diagnostic test, which includes both a beginning-of-year and end-of-year assessment, and is designed to assess proficiency gains over a school year. Rosetta Stone Foundations is an interactive language learning software that teaches all four language skills, with an emphasis on speaking and listening.

Method

Study Design

Eight public schools in a large urban school district in Arizona participated in the study during the 2017–2018 school year. Random assignment was done at the school level. Prior to assignment, it was established that the demographics were highly similar across the eight schools in the study. For the 2017–2018 school year, 99.9% of the 8,194 students at the eight schools received free or reduced-price lunch, a proxy indicator for socioeconomic status (SES) suggesting broadly similar SES across students (Arizona Department of Education Accountability and Assessment, 2018). Additionally, 91% of the 2017–2018 EL population at the eight schools was classified as ethnic Latino. Overall, 20% of students at the eight schools were classified as ELs, and the distribution of ELs was similar for the treatment and control groups. At the treatment schools, 21% of students were classified as ELs, and 19% were classified as ELs at the control group schools (Arizona Department of Education Accountability and Assessment, 2018). Because of the demographic homogeneity of the school population, it was determined that demographic similarity was achieved a priori and pairwise matching along demographics was unnecessary.

Prior to assigning groups to condition, schools were randomly assigned to two groups. One variable considered for random assignment was school type, as there were four middle schools and four K–8 schools. To create two groups, two schools of each type, that is, two middle schools and two K–8 schools, were assigned to a group, with the goal of creating approximately equal numbers of students in each group. From these two groups, one group of four schools was randomly assigned to the treatment group, and the other group of four schools was assigned to the control group.

Participants

Participants were EL students in sixth through eighth grades. School district staff compiled the final list of eligible participants, with a goal of including all learners who required English language support at the lower (preemergent/emergent/basic on the Arizona English Language Learner Assessment [AZELLA] Scale) and intermediate levels of proficiency. Students were identified as needing English language support based on either the previous year's state assessments or screenings for new students. Low- and intermediate-level students were targeted for the intervention based on the affordances of the software program selected. At all eight schools in the study, all students who were enrolled and in school during the pretesting window and met the inclusion criteria of needing English support were included in the study.

The TELL beginning-of-year assessment provides calculated estimates of students' proficiency aligned to the AZELLA. The five AZELLA proficiency levels are (1) preemergent, (2) emergent, (3) basic, (4) intermediate, and (5) proficient. According to estimates from the TELL's beginning-of-year assessment, which was used to establish

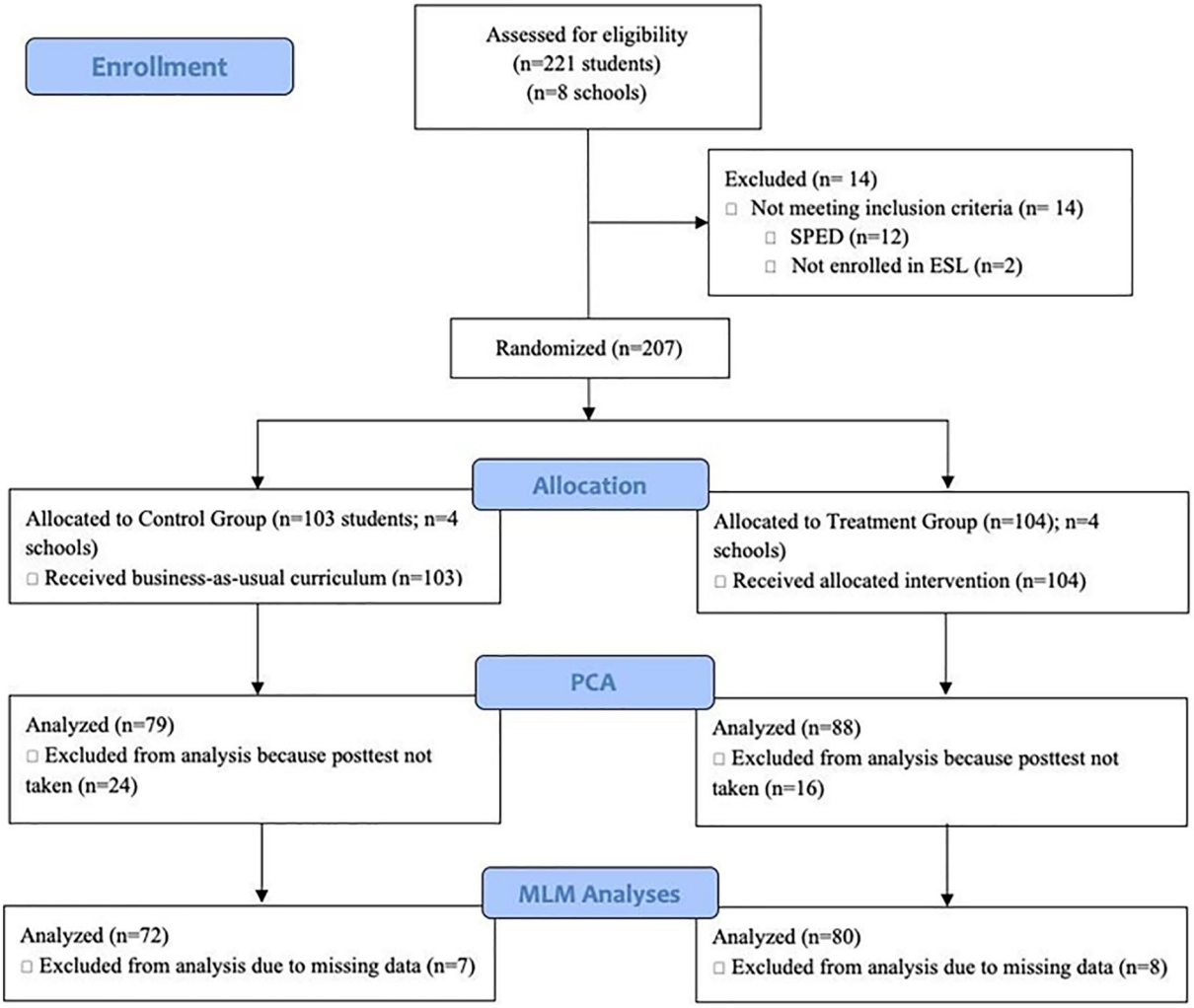

FIGURE 1. *CONSORT flow diagram of the present randomized controlled trial.*
Note. CONSORT = consolidated standards of reporting trials; SPED = special education; ESL = English as a Second Language; PCA = principal components analysis; MLM = multilevel model.

baseline proficiency for the present study, the majority of the students in the study were classified as preemergent-, emergent-, or basic-level ELs in terms of overall proficiency level. Students in the preemergent and emergent levels "lack the English skills to communicate . . ." and "do not demonstrate sufficient skills in English to access mainstream curriculum" (Arizona Department of Education Assessment, 2016). Students in the "basic" level "have a limited understanding of social spoken English" and "respond orally with isolated words and simple sentences with grammatical errors" (Arizona Department of Education Assessment, 2016). TELL results classified 34% of the original sample as intermediate at pretest, and the remaining 6% were classified as proficient overall. The students who were classified as proficient on the TELL were still enrolled in the ESL program because of the official AZELLA placement done by the district in the previous school year. At the intermediate level, students are "limited in their understanding of academic English" and "generally respond orally with phrases and simple sentences" (Arizona Department of Education Assessment, 2016).

At the beginning of the school year, 221 (Control = 110, Treatment = 111) students were pretested (Figure 1). Because 12 of these students (Control = 7, Treatment = 5) had special education status, they were excluded from the final data set due to the small number and therefore lack of ability to draw meaningful conclusions for that population. Two additional students from the Treatment group were excluded because they were not enrolled in an ESL class and did not receive the intervention. Because special education status and ESL class enrollment were student characteristics determined prior to the study, these students were excluded from the total number of students for attrition calculations, per the What Works Clearinghouse guidelines (WWC;

TABLE 1
Number of Students Included in Analysis, by Grade, Gender, and Condition

	Grade 6		Grade 7		Grade 8		
Students	Treatment	Control	Treatment	Control	Treatment	Control	Total
Male	25	16	14	6	15	14	90
Female	9	17	10	10	7	9	62
Total	34	33	24	16	22	23	152

2017). Of the 40 students who were not posttested, 24 were in sixth grade, 13 in seventh grade, and 3 in eighth grade. One student's home language was unknown, but for the remaining 39, their home language was Spanish. At posttest, 167 students had at least some data that could contribute to a principal components analysis (PCA). Due to missingness among TELL measures ($n = 10$) and covariate data ($n = 5$), 152 students (Control = 72, Treatment = 80) were included in the final multilevel statistical models. Overall attrition, with 207 students at pretest and 152 at posttest, was 26.6%; the differential attrition rate was 7% (Control = 30.1%, Treatment = 23.1%). Given that this attrition rate is considered both low under optimistic assumptions and high under conservative assumptions, we calculated baseline equivalence for the final sample. We found the final sample to still be balanced at pretest on overall TELL scores as the calculated effect size of .008 is less than .05 (Control *M*: 429.000, standard deviation [*SD*]: 15.15; Treatment *M*: 428.875, *SD*: 15.79), satisfying baseline equivalence per WWC standards. Additionally, it should be noted that for populations like the one studied here—low-income minority students in an urban school district—greater student mobility is the norm (Welsh, 2017), and increased rates of attrition should be expected.

In the final study sample, 98% of students identified their home language as Spanish (Control = 98.6%, Treatment = 97.5%), and all students received free or reduced-price lunch. The final sample had on average an 89.8% attendance rate, split similarly across the control group (89.0%) and the treatment group (90.5%). Although the entire sample had slightly more males (59%) and sixth graders (44%), students were demographically similar across groups (see Table 1).

Materials

Foundations. Rosetta Stone Foundations English is a self-paced software course intended to supplement teacher-led instruction in Grades K–12 within a blended learning environment. The program uses a target-language-only structure to simulate an immersion environment (Rosetta Stone, 2010). This feature allows for learners (1) with no English and (2) from a variety of native language backgrounds to use the program within a classroom environment where teachers may not be able to communicate in the learners' own languages.

The Rosetta Stone Foundations English course is appropriate for learners who are complete beginners as well as learners who are working on intermediate English content. This design feature informed participant selection in the current study (see the Participants section). The software program is designed to introduce the target language in a highly structured sequence that aims to ensure that students are working at a level just at or slightly beyond their existing skill level, to reduce anxiety and provide many opportunities for success in the new language (Rosetta Stone, 2010). Although all four skills are covered, the curriculum is weighted toward oral-aural language skills including extensive opportunity for speaking and listening practice to build communicative ability and confidence. Students work on the program individually and use headphones and a microphone to interact with the computer for speaking and listening activities. Feedback on speaking accuracy is provided by a built-in speech recognition engine (SRE). The speech models used in the SRE are trained on native speech using machine learning algorithms and evaluate learners' pronunciation of words and phrases. The learners are given a form of yes/no feedback. If learners do not achieve an acceptable score as determined by the SRE, they are prompted to attempt the word or phrase again.

The course comprises 20 units, and each of these units includes a set of four related (topical) lessons that introduce new material and then provide practice opportunities for the student in speaking, listening, reading, writing, needed vocabulary, and targeted grammar (see Figure 2).

Table 2[1] provides an overview of concepts covered in the first eight units, which were the most frequently completed by students in the study. Unit 1 introduces basic vocabulary for people and everyday items, adjectives, colors, and greetings and focuses on the grammar involved in forming plurals, correctly using pronouns, using the present progressive, and forming yes/no questions and question words. As students progress through the program into higher units, they encounter more academic vocabulary that will allow them to connect ideas and organize texts. For example, in Unit 3 they practice question formation with "why" and appropriate responses with the subordinating conjunction "because." Comparative and superlative structures are introduced in Unit 4. Many students in the current study managed to

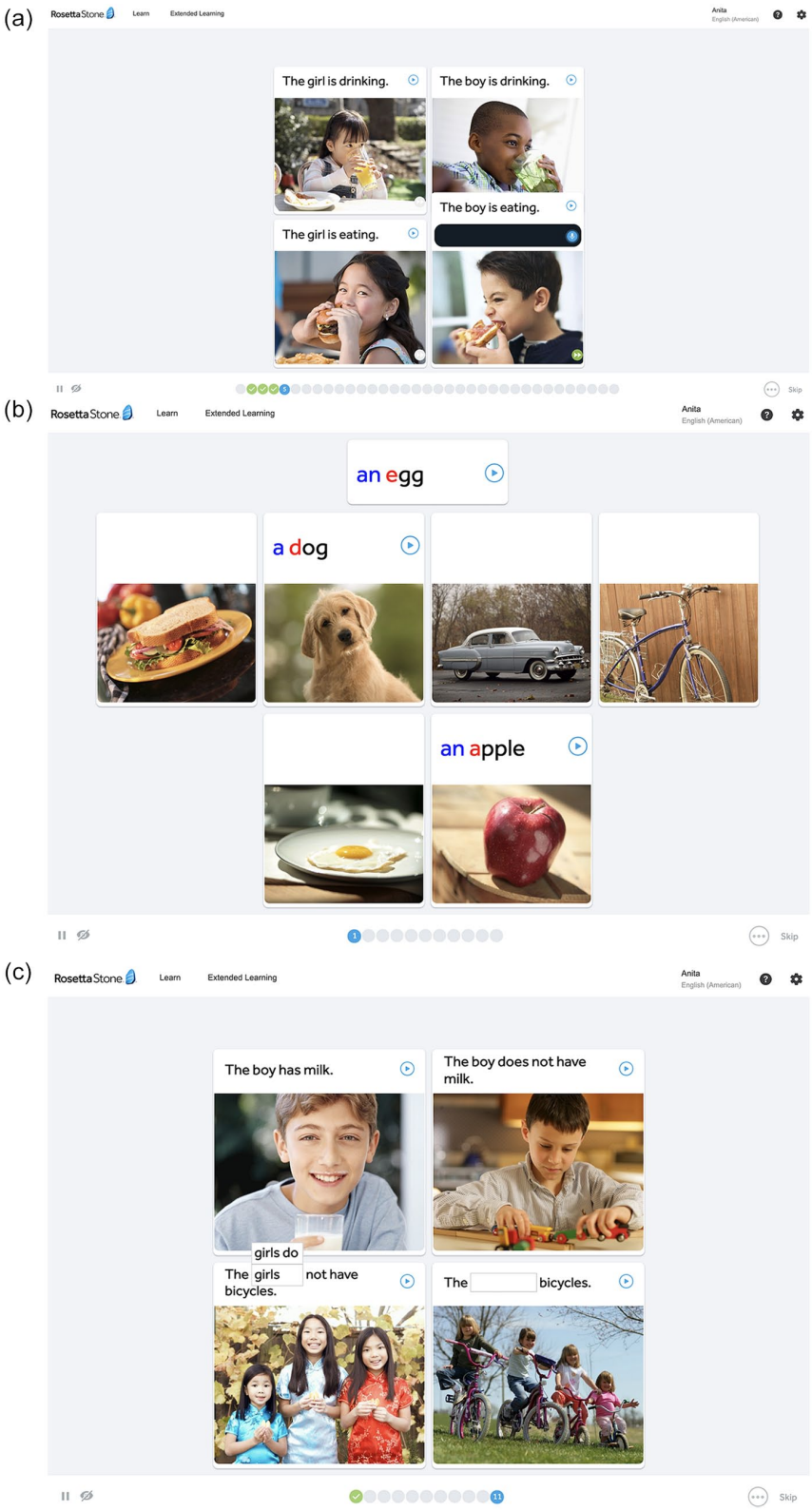

FIGURE 2. *Example screens from the Rosetta Stone Foundations K–12 course.*
Note. (a) A speaking screen. The student is prompted to produce the sentence, "The boy is eating." Feedback is then provided by the Speech Recognition Engine. (b) On this grammar screen, colors are used to highlight important features of English grammar and spelling while introducing basic vocabulary. (c) On this grammar screen, students must demonstrate learning by selecting the correct form or forms from a drop-down menu.

TABLE 2
Student Activity and Sample of Concepts Covered

Unit	Sample of concepts covered	Total number of students introduced to all concepts in each of the four lessons	Total number of students with activity in any or all the four lessons	Average time in hours spent by each active learner
1	Plurals, present progressive, negation, yes/no questions	79 (99%)	80 (100%)	5.2
2	Question words, family relationships, adjectives	64 (80%)	78 (98%)	5.0
3	Simple present, numbers, question formation, "because"	50 (63%)	70 (88%)	5.2
4	Compound sentences, comparatives and superlatives	43 (54%)	60 (75%)	4.9
5	Ordinal numbers, contractions, future tense	33 (41%)	46 (58%)	5.2
6	Past tense, indirect object pronouns, school subjects	19 (24%)	32 (40%)	4.6
7	Comparisons, demonstratives, politeness	14 (18%)	27 (34%)	3.8
8	Negation, sequencing events, modal verbs, emotions	20 (25%)	32 (40%)	4.3

complete Unit 5, where ordinal numbers and sequencing are introduced and practiced.

Table 2 also includes student activity in the program by the final pool of 80 students in the treatment group. For this study, which focused on students with relatively low levels of English proficiency, teachers were asked to have students work for 90 minutes a week on the program, with the goal of completing at least four units of the course. All learners started in Unit 1, and all but one student progressed to Unit 2. The majority of students proceeded linearly within the program and the majority of work was done in Units 1 to 8; however, both teachers and students themselves could advance to different units at their discretion, and 36 of the 80 students explored Units 9 to 20, spending an average of 4 hours in these units. On average, students spent 28 hours in the program and completed the equivalent of five units. A majority of students completed each of the first four units. The expected time needed to complete a unit is 5.6 hours, but students generally spent slightly less time than that on average in each unit.

Over the course of the 2017–2018 school year, there were 32 weeks of activity in the Rosetta Stone program. The first 2 weeks had relatively less usage as students and educators incorporated the program into the curriculum. Over the following 22 weeks, learners averaged the recommended 90 minutes per week. Usage slowed after AZELLA testing and spring break as students used the program for approximately 45 minutes per week on average until the end of the school year.

Test of English Language Learning. The primary measurement tool was the TELL diagnostic test from Pearson (Bonk, 2016). The test is administered on tablet computers and includes both a beginning-of-year test (pretest) and end-of-year diagnostic test (posttest) to measure baseline proficiency and calculate growth in a number of domains. For the sixth- to eighth-grade band, the TELL calculates an overall score, four domain scores, and six subskill scores (see Table 3).

TABLE 3
Test of English Language Learning Domain and Subskill Scores

Domain scores	Subskill scores
Listening	Grammar
Speaking	Vocabulary
Reading	Pronunciation
Writing	Fluency
	Reading rate
	Expressiveness

Because all students in the study were in a single grade band (sixth to eighth), their scores could be compared directly (Bonk, 2016).

The TELL has 22 item types that require either short or extended constructed responses. Many of the item types test multiple language skills simultaneously; the "listen and retell" item type, for example, requires students to listen to audio and then retell it orally in their own words. The "read and summarize" item type requires students to read a passage and then summarize it in writing on the next screen (for more details on the item types and test structure, see Bonk, 2016). As described in Bonk (2016), to establish the

reliability of the TELL, test developers used two methods. In their alternate forms test-retest reliability study, 127 students in the grade band studied here took a version of the TELL two or more times within a week. Correlations were .70 or greater for the four language skill domain scores and the overall score, indicating good reliability: Overall = .87, Listening = .80, Speaking = .77, Reading = .78, and Writing = .70. Correlations derived from the split-half method (Brown, 1996) showed similarly high reliability estimates for the domain scores: Listening = .88, Speaking = .94, Reading = .86, and Writing = .79. The automated scoring of items depends on modality and often relies on existing models such as the latent semantic analysis model developed by Landauer et al. (2003). In the case of speech, scoring is based on features outlined in Bernstein et al. (2010). To ensure that machine-generated scores matched human judgments, 150 field-test takers in each grade band were held out to be scored by human raters and compared with automatically generated scores. Correlations for the sixth- to eighth-grade band were quite high: Overall = .90, Listening = .86, Speaking = .77, Reading = .91, and Writing = .96 (Bonk, 2016).

Using the approach from Bailey et al. (2007), external studies have confirmed that the TELL aligns closely with English Language Development Standards in Arizona (Stevens et al., 2015b), California (Stevens et al., 2015a), Texas (Frantz & Bailey, 2016), and the World-Class Instructional Design and Assessment (WIDA) Consortium (Stevens, 2015), which is composed of 41 U.S. states, territories, and federal agencies. All the alignment studies investigated specific aspects of linguistic forms and language functions and their proportional representations on both tests. In WIDA, for example, the authors found a "high degree of match" between "key functions that are highly represented in WIDA such as Identify, Interpret & Comprehend, and Sequence" (Stevens, 2015). In California, the authors found comparable proportional representations for 88% of linguistic forms and 88% of language functions (Stevens et al., 2015a).

Procedure

In Arizona, in the 2017–2018 school year, ELs received 4 hours of daily instruction that were divided into four 1-hour blocks: (1) oral English conversation and vocabulary instruction, (2) grammar instruction, (3) reading instruction, and (4) writing instruction. For the treatment group, the software was incorporated into the oral English conversation portion of the state's English language proficiency requirements. Control students continued with the district's standard English curriculum, which consisted of vocabulary development protocols such as those by Marzano and Pickering (2005) and Frayer et al. (1969), as well as exercises inspired by Kagan and Kagan (2009). Achieve 3000 (https://www.achieve3000.com), a software program focusing on literacy, was also commonly used in both treatment and control classrooms. In an end-of-year survey of teachers and paraprofessionals at all school sites, 6/8 (75%) respondents from the control group mentioned Achieve 3000 as an additional tool that they used, and 14/18 (78%) from the treatment group mentioned this software intervention.

Taking into account that treatment and control groups had nearly identical attendance rates and received the same amount of daily ESL instruction, we believe that the students received equal amounts of ESL instruction over the course of the year. And while conditions from classroom to classroom can never be identical, district guidance around the curriculum and the equitable provision of tools such as Achieve 3000 would indicate broad uniformity in classroom conditions. The primary difference between groups is the usage of Rosetta Stone Foundations in the treatment group's oral English conversation block.

Prior to the start of the school year, teachers from the treatment school sites participated in an implementation training on how to effectively incorporate the software into the curriculum. Teachers were shown the various activities included in the software and how to access progress reports and supplementary materials. A second training was conducted at the school year's midway point to review implementation and reporting tools. At the first training, a guideline of 90 minutes of software usage per week was set, with the goal being to complete approximately four units in the program. As described above, most classrooms met the recommended target of 90 minutes per week for 22 weeks out of the school year. In the end-of-year survey conducted with teachers and paraprofessionals, 5/18 respondents indicated that they used the administrative reporting tools, and 2/18 reported using the supplemental materials at some point during the school year. Low usage of the supplemental materials was likely due to an already full curriculum, and low adoption of the reporting tools is not atypical in the first year of implementation (see, e.g., Wayman et al., 2017).

District personnel recruited retired teachers to administer both the pretest and posttest. The retired teachers were blind to group assignment. Pretesting was conducted in late August and early September 2017, while posttesting was completed in early May 2018. Both pretesting and posttesting lasted approximately 1 week. During testing, up to six students were tested simultaneously on iPads while being monitored by a proctor. The test is designed to be self-paced, and students were spaced according to Pearson's implementation recommendations to ensure minimal auditory interference for the speaking portions of the test. Students were not informed that the purpose of the test was for a study. Scores were automatically calculated by Pearson's algorithms without hand-scoring.

TABLE 4
Devices and Classroom Setup

School	Devices	Classroom setup
1	Desktop computers	Computer lab
2	Chromebooks	"Split" classroom
3	MacBook Airs (sixth)/iPads (seventh & eighth)	Classroom (sixth)/Computer lab (seventh & eighth)
4	iPads	Computer lab

Researchers monitored usage throughout the school year and responded to requests from teachers, paraprofessionals, or instructional support specialists. The most common requests involved providing licenses to new students, tips on motivating students, or answering technical questions. These interactions were similar to those carried out by a client manager for a normal client.

To assess fidelity of usage and implementation at the treatment schools, classroom observations were conducted in December 2017 by one of the test proctors. The proctor visited all 10 of the treatment classes and documented students' level of engagement with the software, use of the target language in the classroom, classroom setup, and general implementation. At 5-minute intervals, the observer recorded students' level of engagement in the learning activity on a 3-point scale: low engagement (more than half of students engaged), medium engagement (less than half engaged), or high engagement (almost all students engaged). The observer recorded 95% high engagement over the course of the observations. The observer also rated students' and teachers' use of the target language, English, during the session on a 5-point scale (*never*, *seldom*, *about half the time*, *usually*, and *always*). Teachers or paraprofessionals used English always ($n = 5$) or usually ($n = 5$), while students used English usually ($n = 9$) and in one case about half the time. The observations also confirmed that students used four different devices to interact with the software (see Table 4). In most cases, students used the software in a dedicated computer lab; however, at School 2, a single teacher monitored students' Rosetta Stone usage while more advanced ELs (whose data are not included in this study) worked on other tasks in the same classroom. At School 3, some students used laptop computers in the classroom.

Demographic and other relevant variables were obtained directly from the school district at the end of the school year. The list of variables received from the school is as follows:

- Enrollment date
- Grade (6, 7, 8)
- Special education status (yes/no)
- Gender (male/female)
- Home language
- Free/reduced-price lunch status
- Race
- Attendance

Analysis

The collected TELL outcome scores were missing for 73 Speaking scores—38 from 2017 pretesting and 35 in 2018 posttesting—due to technical issues in testing (e.g., participants talked too quietly for the recording to be scored, microphone issues, etc.). This in turn resulted in missing overall TELL scores due to the missing Speaking scores for these students. Fluency, Pronunciation, Reading Rate, and Expressiveness subscale scores were based at least in part on the Speaking scores as well and were likewise missing for these participants.

The initial data set used for a principal components analysis included 167 participants with (across pretest and posttest) 334 scores for Listening, Reading, Writing, Grammar, and Vocabulary variables, and (due to missing data) 261 scores for Speaking, Fluency, Pronunciation, Reading Rate, and Expressiveness variables. Descriptives for these variables are presented in Table 5 later in the text.

The final data set used for multilevel modeling contained 152 participants, having excluded 10 participants for missing composite scores from the PCA and identified covariates of interest including percentage attendance (missing for five additional students), grade level, and sex. All analyses were conducted in R Version 3.4.4 (R Core Team, 2015), using the psych package Version 1.8.3.3 (Revelle, 2018) for PCA and the lme4 package Version 1.1-17 (Bates et al., 2015) for multilevel modeling.

The TELL diagnostic is a norm-referenced test that provides 11 outcome measures. With the exception of reading rate, all scores in the Grades 6 to 8 band range from 400 to 500. Table 6 provides the score ranges and relevant proficiency-level descriptors for the TELL and the corresponding AZELLA alignments derived from a concordance study (Pearson, 2015).

The AZELLA ranges in Table 6 only refer to the overall score, whereas the TELL ranges apply to all domains and subdomains except reading rate. In the case of reading rate, the score reflects actual words correct per minute with some scoring adjustments for errors. As can be seen in Table 5,

TABLE 5
Means (and Standard Deviations) for the 11 Test of English Language Learning Outcome Measures

	RS nonusers			RS users		
Measure	Pretest	Posttest	Gains	Pretest	Posttest	Gains
Overall	428.48 (16.14)	441.62 (17.96)	13.14	429.69 (17.57)	444.72 (17.87)	15.03
Speaking	440.89 (22.98)	449.30 (25.32)	8.41	442.13 (25.54)	455.21 (23.22)	13.08
Listening	437.27 (20.82)	442.54 (20.21)	5.27	438.20 (22.03)	449.62 (21.85)	11.42
Reading	413.96 (14.35)	433.19 (19.21)	19.23	416.24 (13.80)	436.27 (19.59)	20.03
Writing	419.87 (19.62)	430.95 (23.49)	11.08	418.43 (17.85)	430.73 (22.34)	12.30
Grammar	434.32 (13.64)	450.15 (18.72)	15.83	435.58 (15.58)	454.61 (17.96)	19.03
Vocabulary	428.47 (17.79)	437.38 (20.27)	8.91	429.69 (18.41)	440.93 (19.53)	11.24
Fluency	443.65 (23.68)	447.16 (20.56)	3.51	446.80 (26.41)	455.35 (23.44)	8.55
Pronunciation	438.08 (23.81)	442.44 (18.12)	4.36	442.84 (24.61)	453.39 (21.88)	10.55
Reading rate	88.30 (33.55)	95.28 (36.46)	6.98	80.40 (29.41)	99.19 (28.06)	18.79
Expressiveness	421.86 (22.89)	422.77 (25.16)	0.91	420.14 (21.19)	430.03 (23.87)	9.89

Note. RS = Rosetta Stone.

TABLE 6
TELL and AZELLA Score Ranges

TELL		AZELLA	
Score range	Proficiency level	Score range	Proficiency level
400–419	Limited	400–409	Preemergent/Emergent
420–439	Basic	410–432	Basic
440–459	Intermediate	433–452	Intermediate
460–479	High	453–500	Proficient
480–500	Advanced		

Note. TELL = Test of English Language Learning.

scores at pretest for the control and treatment groups were highly similar.

Principal Components Analysis

The TELL outcomes consisted of 11 separate scores, including an overall score, four domain scores, and six subscale scores. Thus, PCA was chosen to reduce the number of dependent variables submitted to the substantive analysis. This was done to protect against type I error (false-positive results), which increases as the number of analyses is increased. PCA is a data reduction technique used to reduce many variables (based on their intercorrelations) into fewer linear composites (called "components") that account for a majority of the proportion of the variance in the original variables. We then interpret these components by examining which variables load onto which components, with an interpretable loading usually defined as a .30 or higher loading onto a component. For example, if "Speaking" and "Listening" scores load heavily on one component, that component could be interpreted as an "Auditory Composite" or a "Speaking-Listening Composite." The more highly a variable loads on a component, the tighter the correspondence should be between increases in a component score and increases in the original variable.

There are various ways of selecting the number of components to calculate from a PCA (e.g., Kaiser's rule, scree plot, percentage of variance explained) but the exploratory nature of PCA leaves it to the researcher to select the number of components most relevant given the data. Oblimin rotation of the components was selected for this analysis, as it allows the resulting components to correlate with one another, which is to be expected from 11 proficiency measures from the same test (e.g., one might expect at least a small correlation between Listening Proficiency and Reading Proficiency measures). With respect to the missing Speaking score data and related scores, we employed a pairwise deletion strategy because PCA with pairwise deletion has been shown to be robust to missing data at the levels observed in this data set (Van Ginkel et al., 2014). To preview the results, we fit PCAs with two, three, and four components, and then selected the model with the clearest interpretation based on

TABLE 7
The Pattern Matrix of Standardized Loadings for the Final Principal Component Analysis Solution

Measure	Component 1	Component 2	Component 3	h^2	u^2
Speaking	**.93**	.09	−.03	.92	.08
Listening	**.95**	.04	−.04	.92	.09
Reading	.19	**.60**	**.33**	.77	.23
Writing	.00	**.93**	.02	.88	.12
Grammar	**.70**	**.36**	.04	.88	.12
Vocabulary	**.72**	**.38**	−.02	.92	.08
Fluency	**.99**	−.20	.06	.86	.14
Pronunciation	**.91**	−.12	.06	.77	.23
Reading rate	.14	.17	**.65**	.63	.37
Expressiveness	−.06	−.04	**.95**	.84	.16

Note. Bold typeface indicates loading > |.3| which indicates at least a moderate loading on the component. h^2 = communality, variance that is shared with other variables; u^2 = uniqueness, variance that is unique to a variable and not shared with other variables.

TABLE 8
Loadings and Proportion of Variance Explained of the Original Variables

Variable	Component 1	Component 2	Component 3
Sums of squared loadings	4.91	1.90	1.58
Proportion variance	.49	.19	.16
Cumulative variance	.49	.68	.84

component loadings. It was discovered that including the overall TELL score in the PCA produced untrustworthy model results (due to a Heywood case, i.e., communality equal to 1) for three- and four-component models; therefore, overall TELL score was dropped and PCAs were rerun with the reduced set of 10 TELL outcomes. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (MSA) on this set of 10 TELL outcomes resulted in an acceptable overall MSA of .79 (Kaiser, 1974). The KMO MSA for each individual scale varied, but all were above .50, indicating acceptable KMO MSA values: Speaking = .69, Listening = .87, Reading = .86, Writing = .55, Grammar = .89, Vocabulary = .73, Fluency = .88, Pronunciation = .88, Reading Rate = .94, and Expressiveness = .77.

PCA Results

After testing two-, three-, and four-component models, the three-component PCA was selected based on the lack of interpretability of the two-component solution and because the four-component solution simply splits the variables that loaded on Component 3 into two less meaningful components. The three-component solution is also satisfactory given an examination of the scree plot and given that the three components explain a vast majority, 84%, of the variance of the original 10 variables. The pattern matrix is presented in Table 7, proportion of variance is explained in Table 8, and the component correlations are presented in Table 9. Variables with loadings greater than or equal to .30 are interpreted as loading on a particular component. The larger the loading, the more heavily a variable influences a particular component. Note that each component has at least one variable loading above .90, suggesting that increases in these components correspond highly to increases in the loading variables. (See the appendix for correlation plots of each component with the original TELL measures.)

TABLE 9
Component Correlation Matrix Resulting From Oblimin Rotation

	Component 1	Component 2	Component 3
Component 1	—	.47	.34
Component 2		—	.31
Component 3			—

Component 1 (C1) consists most heavily of the Speaking, Listening, Grammar, Vocabulary, Fluency, and Pronunciation variables. Given that all these variables involve speaking and/or listening measures, we label C1 the Speaking-Listening TELL Composite.

Component 2 (C2) consists most heavily of the Reading and Writing variables, and, to a lesser degree than for C1, the Grammar and Vocabulary variables. That the latter two variables load on both C1 and C2 is a result of the Grammar and

Vocabulary domains being assessed in both the spoken and written modalities. Given that all of these variables consist of reading and/or writing measures, we label C2 the Reading-Writing TELL Composite.

Component 3 (C3) consists most heavily of the Reading Rate and Expressiveness variables, and, to a lesser degree than for C2, the Reading variable. This latter variable loading on both C2 and C3 is reasonable given that it predominantly represents reading skills—hence the larger loading on C2—with part of the reading skill assessment being done by the student smoothly and accurately reading texts out loud—hence the smaller loading on C3 (Bonk, 2016). Given that all these TELL variables measure, to some extent, the student's ability to make sound-form connections, we label C3 the Reading-Aloud TELL Composite.

Linear Multilevel Modeling

Three separate linear multilevel models (MLMs) were run using the resulting PCA composites as dependent variables. It is important to note for the interpretation of model estimates that PCA composites approximate a mean of 0 and a standard deviation of 1, and thus the estimates for all independent variables indicate predicted changes in the number of standard deviations of the composite, holding all other variables constant. Predictors of interest included test time (pretest vs. posttest; dummy-coded with pretest as baseline), Rosetta Stone software usage (continuous predictor, log-transformed), and their interaction. Covariates included grade level (centered at seventh grade), attendance (percentage, *z*-scored), and sex (simple coded, −.5 female, +.5 male). These covariates were not of interest to this study but were included in all models to control for any possible imbalance or influence on performance. Because SES, race, and home language exhibited minimal variation across students, these variables were excluded from modeling. Rosetta Stone usage (hereafter "RS usage") was operationalized via the number of unique prompts completed within the Foundations software, such that all nonusers have a value of 0 and RS users have a value greater than zero (observed range: 402–15,065). A prompt is defined as any item that can generate a response value, either through writing, speaking, clicking, or tapping. This RS usage variable was log-transformed prior to analysis due to a positively skewed distribution. Unique prompts were used as the RS usage variable instead of total hours recorded with the software because they are a more accurate metric of the amount of content consumed by the user. Number of hours is somewhat less accurate due to administration reasons (e.g., underestimated because of how different devices record usage; overestimated due to participants starting a module and then leaving the computer to use the restroom, etc.).

Given our a priori research hypotheses focused on test time and RS usage, model fixed effects were forced entry, including test time, RS usage, the interaction of test time and RS usage, and simple effects for the covariates grade level, sex, and attendance. With this model parameterization, fixed-effect model parameters reported in the tables below are interpreted as follows. The intercept estimates the expected outcome value on the dependent variable for a nonuser (i.e., someone with zero completed prompts) at pretest (the baseline level for test time) for a student in Grade 7, average attendance, and regardless of sex. Test time estimates the difference between pretest and posttest for the nonuser group (i.e., pre-post gains); a significant effect of test time would indicate that the nonusers showed a significant improvement from pretest to posttest. Because pretest is the baseline level for test time for all models, the variable of RS usage estimates differences in the pretest outcome value for RS users relative to nonusers; a significant positive effect of RS usage would indicate that individuals with greater RS usage had higher pretest scores relative to the nonuser group. And the interaction of test time and RS usage estimates the differences in pre-post gains between the RS nonuser and RS user groups: a nonsignificant interaction would mean the RS users showed gains similar to the nonusers; a significant, positive interaction would indicate that RS usage led to significantly greater gains relative to the nonuser control group. Thus, the interaction term is the most direct test of the hypothesis that RS usage should positively impact learning outcomes.

The random effects structure accounted for the repeated measures (i.e., multiple test scores per subject), and the fact that subjects were nested within teachers nested within schools. In cases where the level explained no variance in the model, the random effects structure was simplified by removing that level. Random slopes for test time (varying by subject, teacher, and/or school) were forward-tested with likelihood ratio tests to arrive at the maximal random effects structure that could be supported by the data (Baayen, 2008; Baayen et al., 2008). Reported models were fit with restricted maximum likelihood estimation to further reduce type I error.

Due to the ongoing debate in calculating *p* values for linear MLMs, only *t* values are provided in lme4 output, so $|t| > 1.65$ is considered marginal ($p < .10$), and $|t| > 2.00$ is considered significant at $p < .05$ (Gelman & Hill, 2007).

Linear Multilevel Modeling Results

The results of the analysis of the Speaking-Listening TELL Composite (C1) as a dependent variable are presented in Table 10.

The simple effect of RS usage was nonsignificant, indicating that pretest scores for both groups were equivalent ($b = -.001$, standard error [*SE*] = 0.03, $t = -0.02$). Students in the control group improved from pre to post on the Speaking-Listening TELL Composite by about 0.33 standard deviations ($b = .332$, $SE = 0.10$, $t = 3.24$). Critically,

TABLE 10
Multilevel Model of the Speaking-Listening Test of English Language Learning Composite (Component 1)

Fixed effects	Estimate	*SE*	*t* Value
Intercept	−0.284	0.16	−1.77
Test time (post)	0.332	0.10	3.24
RS usage (log)	−0.001	0.03	−0.02
Test time × RS usage	0.034	0.02	2.10
Grade level	−0.048	0.10	−0.50
Sex	0.321	0.15	2.12
Attendance	0.004	0.07	0.06
Random effects	**Variance**	***SD***	
Intercepts | Schools/Teachers/Students	0.617	0.79	
Intercepts | Schools/Teachers	0.029	0.17	
Intercepts | Schools	0.034	0.19	
Residual	0.275	0.52	

Note. RS usage is log-transformed. Grade is centered at seventh grade. Sex is simple coded so the model intercept reflects the mean of cell means and male is positive. Attendance is *z*-scored. RS = Rosetta Stone; *SD* = standard deviation; *SE* = standard error.

for RS users, increased usage was related to greater pre to post improvements ($b = .034$, $SE = 0.02$, $t = 2.10$) above and beyond the improvement found for the control group. This can be seen visually by examining the distance between the pretest and posttest lines in Figure 3. Students who used RS the most (i.e., with log-transformed RS usage values of 9.62, about 15,000 unique prompts) improved by about 0.66 standard deviations (based on model estimates, which can be interpreted as an effect size). In other words, RS users who used Foundations the most increased nearly twice as much on the Speaking-Listening TELL Composite over nonusers, holding all other covariates constant (see Figure 3). RS users who used Foundations the least (i.e., 402 unique prompts, a log-transformed value of about 6.00), increased more than 60% over nonusers on the Speaking-Listening TELL Composite.

As for the covariates, controlling for the other variables in the model, grade level and attendance were not significant predictors of C1; however, sex was significant, such that male students overall scored significantly higher than female students on this Speaking-Listening TELL Composite. Due to the centering of the sex variable in the model, the above results and figure are controlling for this difference and presenting overall effects of the predictors of interest on average. As for the random effects structure, no random slopes were significant and the vast majority of variance in the random effects structure is captured at the student level (by-group intraclass correlation coefficient [ICC]: .646), with a minority of the variance about equally captured by the teacher (ICC = .031) and school (ICC = .036) levels.

The results of the Reading-Writing TELL Composite (C2) as a dependent variable are presented in Table 11.

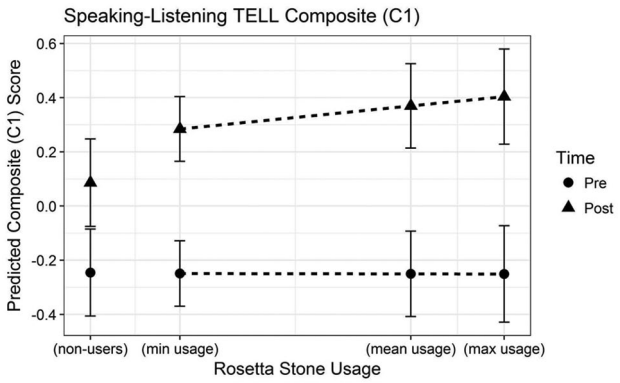

FIGURE 3. *Plot of modeled pretest and posttest scores on the Speaking-Listening Test of English Language Learning Composite.*
Note. Because of a skew in usage, log-transformed usage was modeled, ranging approximately from a minimum usage of 402 unique prompts (log: 6) to a maximum of 15,000 unique prompts (log: 9.62).

The two groups did not show any differences on pretest Reading-Writing TELL Composite scores ($b = -.001$, $SE = 0.03$, $t = -0.03$). The groups also improved similarly from pre to post (see Figure 4) on the Reading-Writing TELL Composite by about 0.92 standard deviations (shown by the effect of test time for the control group: $b = .923$, $SE = 0.11$, $t = 8.13$); the lack of a significant test time × RS usage interaction indicates that RS users showed similar gains to nonusers ($b = -.001$, $SE = 0.02$, $t = -0.04$).

As for the covariates, controlling for the other variables in the model, sex and attendance were not significant predictors of C2; however, grade level was significant, such that students in higher grades scored significantly higher on the

TABLE 11
Multilevel Model to Predict the Reading-Writing Test of English Language Learning Composite (Component 2)

Fixed effects	Estimate	*SE*	*t* Value
Intercept	−0.349	0.16	−2.13
Test time (post)	0.923	0.11	8.13
RS usage (log)	−0.001	0.03	−0.03
Test time × RS usage	−0.001	0.02	−0.04
Grade level	0.179	0.08	2.21
Sex	0.126	0.14	0.91
Attendance	0.010	0.07	0.15
Random effects	**Variance**	***SD***	
Intercepts | Schools/Students	0.459	0.68	
Intercepts | Schools	0.052	0.23	
Residual	0.348	0.59	

Note. RS usage is log-transformed. Grade is centered at seventh grade. Sex is simple coded so the model intercept reflects the mean of cell means and male is positive. Attendance is *z*-scored. RS = Rosetta Stone; *SD* = standard deviation; *SE* = standard error.

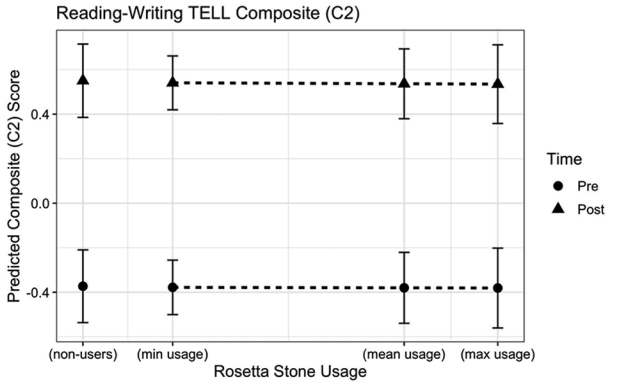

FIGURE 4. *Plot of modeled pretest and posttest scores on the Reading-Writing Test of English Language Learning Composite.*
Note. Because of a skew in usage, log-transformed usage was modeled, ranging approximately from a minimum usage of 402 unique prompts (log: 6) to a maximum of 15,000 unique prompts (log: 9.62).

Reading-Writing TELL Composite. Due to the centering of the grade level variable in the model, the above results and figure are controlling for this difference and presenting overall effects of the predictors of interest on average at Grade 7. As for the random effects structure, no random slopes were significant and the vast majority of variance in the random effects structure is captured at the student level (by-group ICC = .534), with a minority of the variance at the school level (ICC = .061), and zero variance captured by the teacher level, which was removed from the model.

The results of the Reading-Aloud TELL Composite (C3) as a dependent variable are presented in Table 12.

RS users and nonusers were not significantly different on the Reading-Aloud TELL Composite pretest scores (b = −.009, SE = 0.02, t = −0.43). RS nonusers did not significantly improve from pre to post (b = .162, SE = 0.14, t = 1.18). However, RS usage did have a positive effect on pre-to-post improvement (b = .046, SE = 0.02, t = 2.14). Students who used RS the most (again, a log-transformed RS usage of 9.62 or about 15,000 unique prompts in the software) improved by about 0.44 standard deviations on this composite from pre to post over the RS nonusers. In other words, RS users who completed the most unique prompts in Foundations showed more than triple the gains on the Reading-Aloud TELL Composite over nonusers, holding all other covariates constant (see Figure 5). Note that this increase is so large in part because the nonusers did not show a statistically significant improvement pre to post, whereas the RS users did. RS users who used Foundations the least (i.e., 402 unique prompts, a log-transformed value of about 6.00), showed more than a two-and-a-half times greater gain on the Reading-Aloud TELL Composite.

As for the covariates, controlling for the other variables in the model, grade level, sex, and attendance were not significant predictors of C3. As for the random effects structure, no random slopes were significant and the vast majority of variance in the random effects structure is captured at the student level (by-group ICC = .445), with a minority of the variance at the school level (ICC = .004), and zero variance captured by the teacher level, which was removed from the model.

Limitations

As an RCT, this study was designed to provide a strong test of the evidence for the intervention in question. However, as with all research conducted in applied educational settings, challenges related to implementation, participant availability, and available resources led to a number of limitations that might affect interpretation or generalizability of the results.

First, the research team chose to focus on how Rosetta Stone Foundations software might serve a specific population of ELs. By design, a sample of primarily lower level learners in Grades 6 to 8 was selected as a population of learners who might be expected to be well served by the intervention under investigation. Although a larger sample would have been desirable, the intense nature of the four skills pre- and post-testing employed in the study and the focus on lower level middle school ELs necessarily restricted the study sample.

Second, the selection of a partner school district with a relatively homogenous student population in terms of demographics such as first language, SES, and ethnicity—although serving to reduce some of the variance that might be particularly problematic given the study sample

TABLE 12
Multilevel Model to Predict Reading-Aloud Test of English Language Learning Composite (Component 3)

Fixed effects	Estimate	*SE*	*t* Value
Intercept	−0.120	0.13	−0.94
Test time (post)	0.162	0.14	1.18
RS usage (log)	−0.009	0.02	−0.43
Test time × RS usage	0.046	0.02	2.14
Grade level	0.074	0.08	0.88
Sex	−0.183	0.14	−1.26
Attendance	−0.009	0.07	−0.12
Random effects	**Variance**	***SD***	
Intercepts | Schools/Students	0.424	0.65	
Intercepts | Schools	0.003	0.06	
Residual	0.525	0.72	

Note. RS usage is log-transformed. Grade is centered at seventh grade. Sex is simple coded so the model intercept reflects the mean of cell means and male is positive. Attendance is *z*-scored. RS = Rosetta Stone. *SD* = standard deviation; *SE* = standard error.

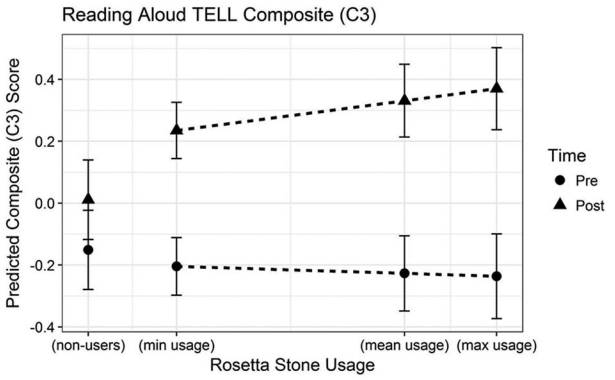

FIGURE 5. *Plot of modeled pretest and posttest scores on the Reading-Aloud Test of English Language Learning Composite.*
Note. Because of a skew in usage, log-transformed usage was modeled, ranging approximately from a minimum usage of 402 unique prompts (log: 6) to a maximum of 15,000 unique prompts (log: 9.62).

size—might also reduce the generalizability of the study's findings. That is, whereas this study finds evidence that the Rosetta Stone Foundations intervention provided an advantage to students in the experimental group, specifically in oral-aural skills, such a finding may not generalize to other age groups, students at other levels of English, or students with different first languages or other background characteristics.

Third, the study sample size was further reduced by two factors outside the control of the researchers: (1) the population under investigation—students of Hispanic origin in a low-income urban school district—can be prone to attrition for reasons unrelated to the study itself (relocation due to shifts in employment location or type; housing instability, etc.) and (2) the necessity of small-group testing with the speaking portion of the TELL sometimes led to poor signal as some students spoke too softly or had technical issues leading to missing data for speaking tasks, reducing the amount of data available for analyses in the final models.

Finally, logistical challenges meant students were assigned to condition by school, rather than by individual, and only the experimental classrooms were observed (to verify fidelity of software usage). The MLMs do account for variation in teacher and school in the random effects structure, and it is worth noting that the vast majority of variance for all of these models is at the student level, with a minority at the teacher and school levels for the Speaking-Listening model, and a minority at the school level and zero variance at the teacher level for the Reading-Writing and Reading-Aloud models. This may suggest not only a lack of meaningful variation at the teacher and school levels (consistency between classrooms and schools) on the TELL but may also suggest a certain uniformity in the implementation of the software across teachers and schools. However, without a detailed analysis of the specific lessons employed in the control classrooms, it is possible that some other difference in instruction might account for these findings.

Discussion

This RCT evaluated the effectiveness of a software intervention, Rosetta Stone Foundations, for ELs in Grades 6 to 8 over the course of one school year. Results indicate that the educational technology intervention improved learning outcomes for these students. Specifically, the intervention contributed to significant improvements in oral/aural skills (i.e., speaking, listening, and reading aloud) when compared with the typical (control) curriculum used by the partner district. In addition to these group-level findings, analyses of individual learning results demonstrated that, within the experimental group, the amount of software usage positively predicted learning gains for these same skills. Together, these results—that the experimental group showed greater gains than the control group and that learners within the experimental group showed greater gains with higher usage of the software—provide compelling evidence that it was, in fact, the software usage that was driving improvement in oral-aural English skills for these low-proficiency learners.

Importantly, for this specific population (low-proficiency ELs), educational technology affords personalized instruction that gives students more opportunities for speaking practice and allows students to proceed at their own pace. Additionally, frequent automated feedback on speaking accuracy allows students to practice speaking without fear of social embarrassment and gauge their own progress. Thus, in this context, technology may provide a low-anxiety

learning environment and lead to a subsequent willingness to participate more in English in the classroom.

This study aligns with existing research that suggests production practice leads to improved retention (Boiteau et al., 2014). Because students were randomly assigned, by school, to condition, and performed equivalently at pretest, this study provides strong evidence, based on Every Student Succeeds Act (2015) criteria, for the effectiveness of Rosetta Stone Foundations for ELs within a blended learning program.

Although the software intervention does train reading and writing skills as well (just not to the same extent as oral and aural skills), and despite others' findings of positive transfer from well-developed oral proficiency leading to improved reading comprehension and writing (August & Shanahan, 2006), no effect of this instructional software was observed for reading and writing skills above and beyond the typical classroom curriculum in this study. This lack of effect could be due to the software curriculum (e.g., perhaps oral skills have to be trained in specific ways to facilitate transfer to another modality), or, perhaps more likely, the students in this study simply did not have a high enough oral proficiency in English to observe a transfer effect. Based on our findings and that of other studies (e.g., August & Shanahan, 2006), we could hypothesize that a language technology intervention with individualized practice at higher levels of starting proficiency has greater potential to contribute to the education and intellectual development of students more broadly.

Conclusion

The purpose of this study was to evaluate the effectiveness of a software intervention that provides individualized practice at a personalized pace and level for middle school ELs. The intervention is unique in its focus on training oral/aural English skills rather than solely literacy skills, and the testing instrument also provides a welcome focus on the full breadth of language skills. Future research should seek to replicate the current study in other states and school districts that have different curricula and less homogenous EL populations. Additionally, further study of software interventions that focus on oral/aural English skills is needed.

ELs in the United States face a range of challenges, and it is critical to investigate what works for this growing population. Hedges (2018) recently underlined the importance of "building usable knowledge" that can impact students' lives. The current article reports the results of an RCT that provides evidence for the type of intervention that can impact thousands of students across the United States and help ELs close the achievement gap more quickly. Hedges also noted that there are many challenges to education research and building usable knowledge. This RCT required many elements to succeed as a study and build that knowledge, notably,

- extensive cooperation, flexibility, and engagement from our participating district;
- collaboration between Rosetta Stone's research team, external researchers, and district staff; and
- a reliable and validated instrument to measure multiple skills.

The study was implemented in a way to ensure ecological validity, and the results provide strong evidence that educational technology can drive positive gains in second language oral and aural proficiency outcomes over even a single academic year. For middle school students in the United States who are nonnative speakers of English, this technology could serve as an important enablement tool to enhance their scholastic achievement.

Appendix

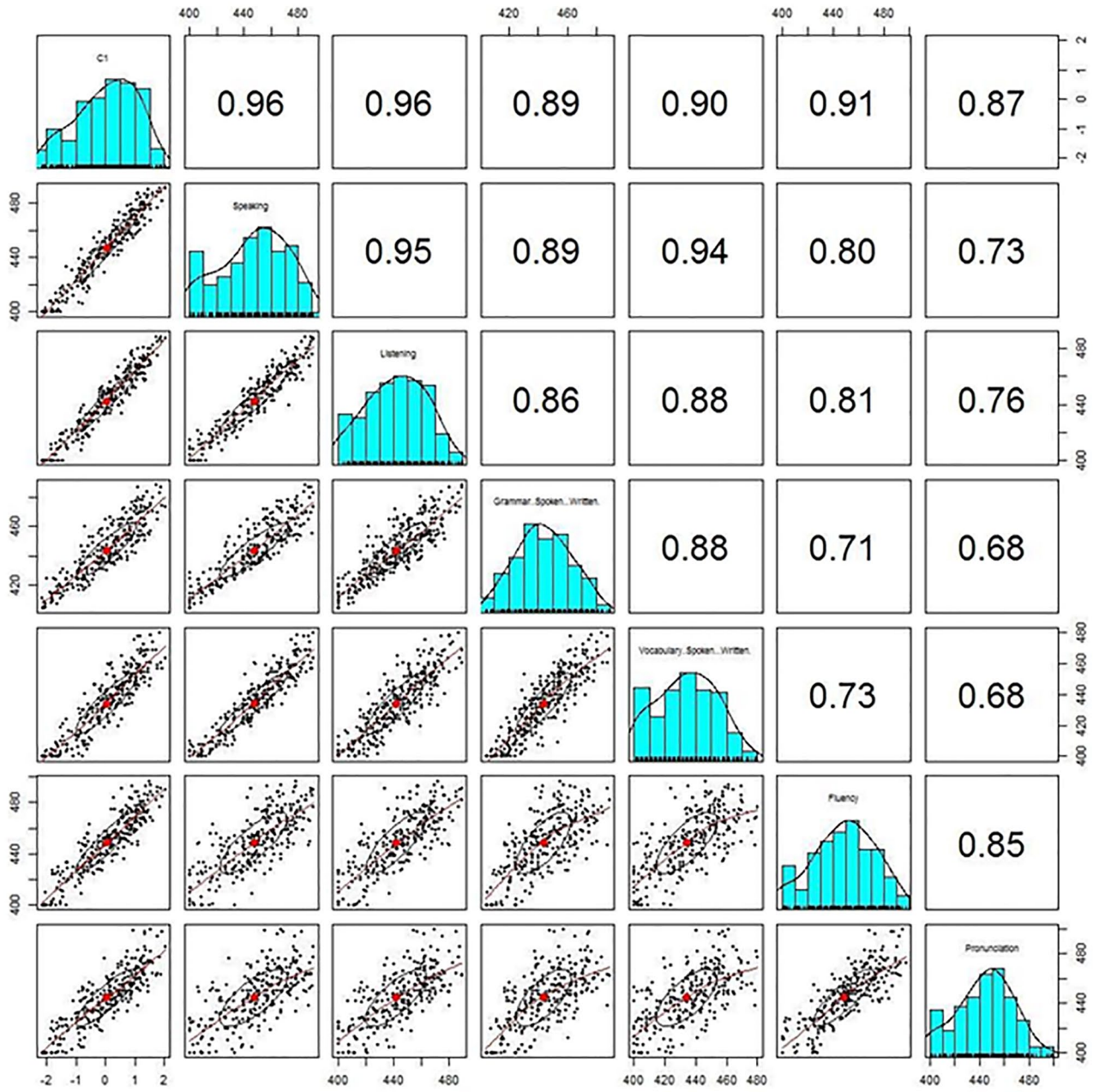

FIGURE A1. *Correlation plots of C1 and its constituent Test of English Language Learning outcomes.*

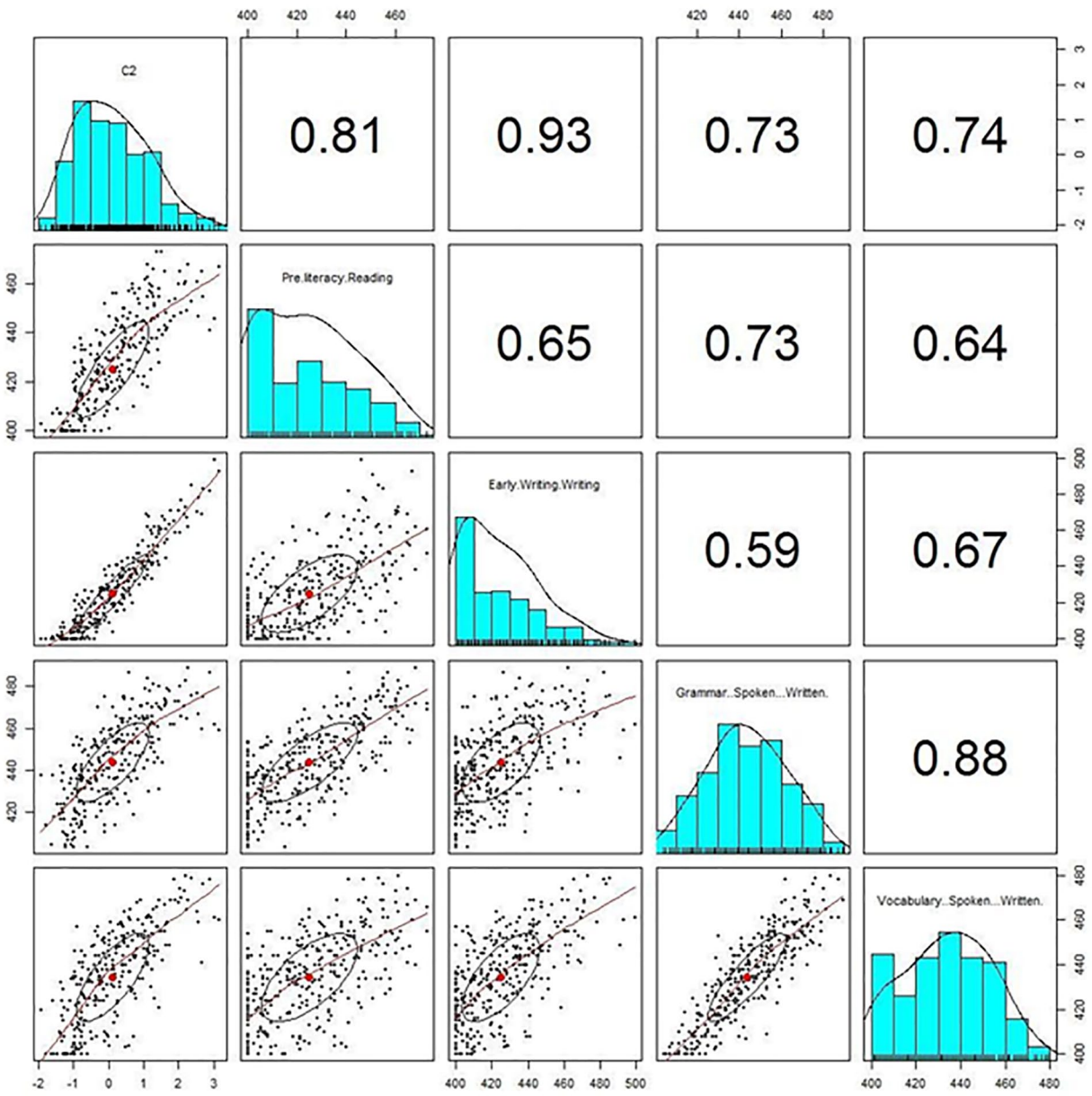

FIGURE A2. *Correlation plots of C2 and its constituent Test of English Language Learning outcomes.*

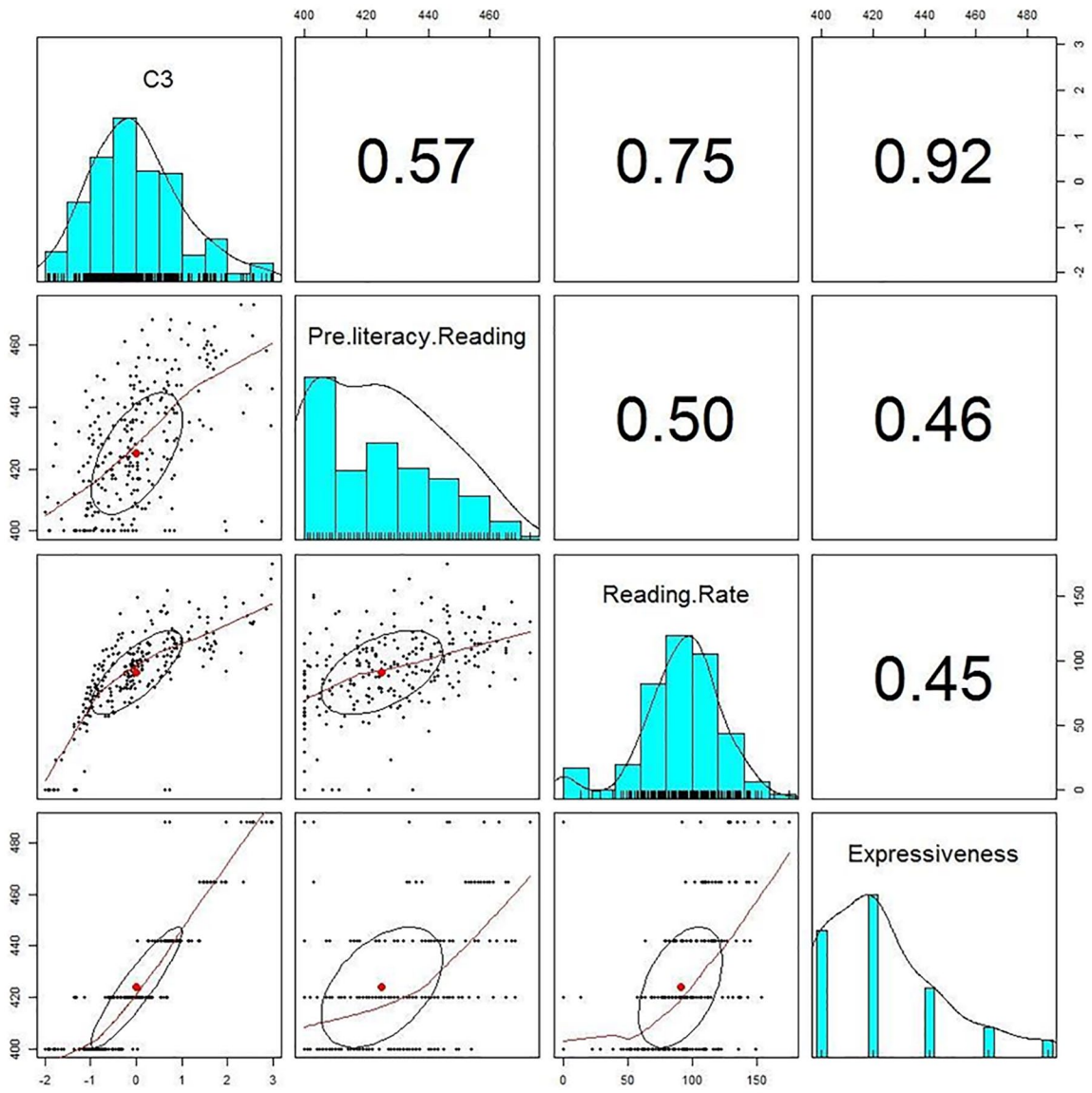

FIGURE A3. *Correlation plots of C3 and its constituent Test of English Language Learning outcomes.*

Acknowledgments

This study would not have been possible without the incredible cooperation of the school district's staff, personnel, and retired teachers, who helped with many aspects of this study. We would also like to thank Dr. Ewa Golonka for her helpful comments on an earlier version of this article. Any errors and mistakes are entirely our own. Three of the authors work for the funding company, Rosetta Stone. The other two authors were hired by the company from the University of Maryland as external consultants for this research project. The external consultants developed the analysis plan, performed the analysis, and were responsible for writing the results. The consultants verified the claims made in the discussion as well as the overall article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research described in this manuscript was funded by Rosetta Stone, Inc.

Note

1. For more information on the full scope and sequence of Rosetta Stone Foundations, see https://resources.rosettastone.com/support/SF/Resources/ScopeAndSequence.pdf

References

Arizona Department of Education Accountability and Assessment. (2018). *2017-2018 School year*. https://www.azed.gov/accountability-research/data/

Arizona Department of Education Assessment, Office of English Language Acquisition Services, and Accountability. (2016). *Guide to navigating and using AZELLA reports*. https://cms.azed.gov/home/GetDocumentFile?id=585073c0aadebe0988f82bda

August, D., & Shanahan, T. (2006). *Developing literacy in second-language learners: Report of the National Literacy Panel on Language Minority Children and Youth*. Lawrence Erlbaum.

Baayen, H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Baayen, H., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bailey, A. L., Butler, F. A., & Sato, E. (2007). Standards-to-standards linkage under Title III: Exploring common language demands in ELD and science standards. *Applied Measurement in Education*, *20*(1), 53–78. https://doi.org/10.1080/08957340709336730

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4. R Package Version 1.1-17*. https://CRAN.Rproject.org/package=lme4

Bernstein, J., van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*(3), 355–377. https://doi.org/10.1177/0265532210364404

Boiteau, T. W., Malone, P. S., Peters, S. A., & Almor, A. (2014). Interference between conversation and a concurrent visuomotor task. *Journal of Experimental Psychology: General*, *143*(1), 295–311. https://doi.org/10.1037/a0031858

Bonk, W. J. (2016). *Scoring TELL (Test of English Language Learning)*. Pearson. https://cdn2.hubspot.net/hubfs/559254/TELL/TELL_White_Paper_4_-_Scoring_TELL.pdf?t=1481317295078

Brown, J. D. (1996). *Testing in language programs*. Prentice-Hall Regents.

Cheung, A. C., & Slavin, R. E. (2012). How features of educational technology applications affect student reading outcomes: A meta-analysis. *Educational Research Review*, *7*(3), 198–215. https://doi.org/10.1016/j.edurev.2012.05.002

Donald, S. (2010). Learning how to speak: Reticence in the ESL classroom. *ARECLS*, *7*, 41–58.

Every Student Succeeds Act, Pub. L. No. 114-95, § 114 Stat. 1177 (2015).

Frantz, R. S., & Bailey, A. L. (2016). *TELL item/item type—State ELD/P standards alignment study report: Texas* [White paper]. Pearson Assessments.

Frayer, D. A., Frederick, W. C., & Klausmeier, H. J. (1969). *A schema for testing the level of concept mastery* [Working Paper No. 16]. Research and Development Center for Cognitive Learning.

Fry, R. (2008). *The role of schools in the English Language Learner achievement gap*. Pew Hispanic Center.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning*, *27*(1), 70–105. https://doi.org/10.1080/09588221.2012.700315

Hedges, L. V. (2018). Challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness*, *11*(1), 1–21. https://doi.org/10.1080/19345747.2017.1375583

Hopman, E. W. M., & MacDonald, M. C. (2018). Production practice during language learning improves comprehension. *Psychological Science*, *29*(6), 961–971. https://doi.org/10.1177/0956797618754486

Kagan, S., & Kagan, M. (2009). *Kagan cooperative learning*. Kagan.

Kaiser, H. F. (1974). An index of factor simplicity. *Psychometrika*, *39*(1), 31–36. https://doi.org/10.1007/BF02291575

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. Shermis, & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary approach* (pp. 87–112). Lawrence Erlbaum.

Marzano, R. J., & Pickering, D. J. (2005). *Building academic vocabulary: Teacher's manual*. Association for Supervision and Curriculum Development.

McFarland, J., Hussar, B., Zhang, J., Wang, X., Wang, K., Hein, S., Diliberti, M., Forrest Cataldi, E., Bullock Mann, F., & Barmer, A. (2019, May). *The condition of education 2019* [NCES 2019-144]. U.S. Department of Education, National Center for Education Statistics. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2019144

National Academies of Sciences, Engineering, and Medicine. (2017). *Promoting the educational success of children and youth learning English: Promising futures*. National Academies Press. https://www.nap.edu/catalog/24677/promoting-the-educational-success-of-children-and-youth-learning-english

Pearson. (2015). *Estimating AZELLA total combined scores from TELL overall scores*. https://www.pearsonassessments.com/

Plonsky, L., & Ziegler, N. (2016). The CALL-SLA interface: Insights from a second-order synthesis. *Language Learning & Technology*, *20*(2), 17–37. https://www.lltjournal.org/item/2945

R Core Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2018). *psych: Procedures for personality and psychological research*. Northwestern University. https://CRAN.R-project.org/package=psych

Richards-Tutor, C., Baker, D. L., Gersten, R., Baker, S. K., & Smith, J. M. (2016). The effectiveness of reading interventions for English learners: A research synthesis. *Exceptional Children*, *82*(2), 144–169. https://doi.org/10.1177/0014402915585483

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rosetta Stone. (2010). *Research basis for the dynamic immersion method* [White paper]. Rosetta Stone.

Stacey, E., & Gerbic, P. (2009). Introduction to blended learning practices. In E. Stacey, & P. Gerbic (Eds.), *Effective blended learning practices: Evidenced-based perspectives in ICT-Facilitated education* (pp. 1–20). Information Science Reference.

Stevens, R. (2015, September). *Report on the alignment between TELL item/item types and the WIDA™ standards* [White paper]. Pearson Assessments.

Stevens, R., Bailey, A. L., & Pitsoulakis, D. (2015a, September). *TELL item/item type—State ELD/P standards alignment study report: California* [White paper]. Pearson Assessments.

Stevens, R., Bailey, A. L., & Pitsoulakis, D. (2015b, November). *TELL item/item type—State ELD/P standards alignment study report: Arizona* [White paper]. Pearson Assessments.

Sutcher, L., Darling-Hammond, L., & Carver-Thomas, D. (2016). *A coming crisis in teaching? Teacher supply, demand, and shortages in the U.S*. Learning Policy Institute.

Troia, G. A. (2004). Migrant students with limited English proficiency: Can Fast ForWord Language™ make a difference in their language skills and academic achievement? *Remedial and Special Education*, *25*(6), 353–356. https://doi.org/10.1177/07419325040250060301

U.S. Department of Education. (2018). *Educator toolkit: Using educational technology—21st Century supports for English learners*. https://tech.ed.gov/edtech-english-learner-toolkits/educators/

Van Ginkel, J. R., Kroonenberg, P. M., & Kiers, H. A. L. (2014). Missing data in principal components analysis of questionnaire data: A comparison of methods. *Journal of Statistical Computation and Simulation*, *84*(11), 2298–2315. https://doi.org/10.1080/00949655.2013.788654

Wayman, J. C., Shaw, S., & Cho, V. (2017). Longitudinal effects of teacher use of a computer data system on student achievement. *AERA Open*, *3*(1), 1–18. https://doi.org/10.1177/2332858416685534

Welsh, R. O. (2017). School hopscotch: A comprehensive review of K-12 student mobility in the United States. *Review of Educational Research*, *87*(3), 475–511. https://doi.org/10.3102/0034654316672068

What Works Clearinghouse. (2017). *What works clearinghouse standards handbook Version 4.0*. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

Authors

DAVID HARPER is a researcher at Rosetta Stone. His research interests include computer-assisted language learning and its applications, second language acquisition, and program evaluation.

ANITA R. BOWLES is head of Academic Research and Learner Studies at Rosetta Stone. Her research focuses on foreign language learning, language aptitude, working memory, tonal languages, and the brain bases of language processing.

LAUREN AMER is a researcher at Rosetta Stone. She is interested in the social impact of learning interventions.

NICK B. PANDŽA is a senior faculty research specialist at the Applied Research Lab for Intelligence & Security and a PhD candidate in second language acquisition at the University of Maryland. His research interests include the effects of individual differences and varied training interventions on language learning and language processing.

JARED A. LINCK was a research scientist at the Applied Research Lab for Intelligence & Security at the University of Maryland. He is now a senior systems engineer at SAS Institute. His research interests include bilingualism, language processing, speech production, executive functioning, and second language acquisition.