jespar
Journal of Education for Students Placed At Risk
Volume 27, Number 1, January–March 2022

Routledge
Taylor & Francis Group

Journal of Education for Students Placed at Risk (JESPAR)

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hjsp20

Evaluating the Efficacy of an English Language Development Program for Middle School English Learners

Erin A. Chaparro, Keith Smolkowski, Barbara Gunn, Caroline Dennis & Patricia Vadasy

To cite this article: Erin A. Chaparro, Keith Smolkowski, Barbara Gunn, Caroline Dennis & Patricia Vadasy (2022): Evaluating the Efficacy of an English Language Development Program for Middle School English Learners, Journal of Education for Students Placed at Risk (JESPAR), DOI: 10.1080/10824669.2022.2045993

To link to this article: https://doi.org/10.1080/10824669.2022.2045993

JOURNAL OF EDUCATION FOR STUDENTS PLACED AT RISK (JESPAR)
https://doi.org/10.1080/10824669.2022.2045993

Evaluating the Efficacy of an English Language Development Program for Middle School English Learners

Erin A. Chaparro[a,b], Keith Smolkowski[a], Barbara Gunn[a], Caroline Dennis[a], and Patricia Vadasy[a]

[a]Oregon Research Institute; [b]University of Oregon

ABSTRACT

This paper reports the outcomes of an experimental evaluation of *Direct Instruction Spoken English* (*DISE*), an English language curriculum that focuses on developing English speaking, listening, and comprehension skills in English Learners (ELs). Twenty-nine middle schools in three states were randomly assigned to teach *DISE* Level 1 or their typical English language development program. Treatment teachers received two days of training and taught daily lessons. Project staff assessed 746 sixth and seventh-grade students with a proximal measure of English language proficiency and distal measures of language and oral reading fluency. Analyses of oral English language suggested differences between intervention conditions favored the *DISE* condition for students who began with lower English language proficiency and accumulated when taught over two years. Differences between intervention conditions were not observed after only one year of instruction and for students with advanced levels of English language proficiency. The findings suggest that an evidence-based English oral language program that included frequent teacher demonstrations and opportunities for students to practice speaking English may improve English oral language skills for middle school students.

A persistent achievement gap exists between the large number of English Learners in U.S. schools and their English-speaking peers. Many of these students enter middle and high school with conversational English but insufficient academic English language skills to perform routine classroom work. This paper describes a study that investigated the efficacy of English oral language instruction for middle school English learners designed to improve English language proficiency and overall academic outcomes.

English learners, a growing population

Emergent bilingual is a broad term used to describe students who are learning to speak a second language. Under the umbrella term of emergent bilinguals, the National Center for Education Statistics (NCES, 2020) defines English learners (ELs) as

CONTACT Barbara Gunn barbarag@ori.org Oregon Research Institute, 1776 Millrace Drive, Eugene, OR 97403, USA.

Supplemental data for this article is available online at https://doi.org/10.1080/10824669.2022.2045993.

individuals who have difficulty speaking, reading, writing, or understanding English. The difficulty must be sufficient to prevent them from learning successfully in classrooms where English is the language of instruction. The number of students federally identified as ELs in U.S. schools, both domestic- and foreign-born, has increased in the last forty years. From 2000 to 2017, the population grew from 8.1% to 10.1% of U.S. students (NCES, 2020). Native Spanish speakers represent the largest and fastest-growing subgroup (Arens et al., 2012), and Hispanic students make up about 25% of K–12 enrollments (Bauman, 2017). In 2017, five million EL students in the K–12 systems, both domestic- and foreign-born, required specialized language learning services (NCES, 2020). Although the umbrella term emergent bilingual is considered more asset-based than English learner, we will refer to the target population as ELs in this paper for clarity and alignment with the federally recognized definitions.

The need to improve outcomes for students identified as ELs

The intransigent gap between ELs and their English-speaking peers, and the resulting challenges, have been widely documented in research and national reports (e.g., Callahan, 2005; Geva & Farnia, 2012; Halle, Hair, Wandner, McNamara, & Chien, 2012; National Academies of Sciences, Engineering, and Medicine [NASEM], 2017; Slama, 2012). While EL students may enter their new schools with a wealth of life experiences and basic conversational English, they may have insufficient academic language skills in English to master subject-area content. The limited academic English puts them at a higher risk of failure on high-stakes tests, high school dropout, and dual classification for special education (Kim, 2011; Umansky, 2016; Umansky, Thompson, & Díaz, 2017). Educators and education scientists have a responsibility to identify and implement effective English language instructional approaches to improve outcomes for EL students (Baker et al., 2014; Saunders, Goldenberg, & Marcelletti, 2013). The number of EL students in U.S. schools is increasing, and they consistently lag behind their peers in academic achievement. In 2017 only 14% of 4th grade EL students were at or above proficient in mathematics, and only 9% were at or above proficient in reading on the National Assessment of Educational Progress (NAEP). In 8th grade, the numbers were even lower (U.S. Department of Education, 2017). These national trends evidence the considerable work that remains on how to teach the English language effectively and improve the overall academic outcomes for ELs.

Further English language curricula research needed

Given the range of languages spoken by students in U.S. schools, effective classroom teaching often requires instruction in English. Schools can teach students English, but to do so, teachers need empirically supported, easy-to-implement instructional tools or curricula. Unfortunately, few English language development (ELD) curricula have been rigorously evaluated under controlled settings with adequate sample sizes to recommend their use in U.S. schools. Fewer have been deemed effective for middle school EL students, and limited evidence is available on the optimal design and delivery

of effective and efficient English language instruction (Arens et al., 2012; Saunders et al., 2013).

Multiple studies have addressed comprehension and vocabulary interventions for EL students, but the programs under study focused on teaching reading and comprehension skills. They did not specifically target the development of English oral language skills. What Works Clearinghouse (WWC, 2012) identified only a few programs that met WWC evidence standards of potentially positive or positive results for EL students in 6th through 8th grade. Borman, Park, and Min (2015) study on *Achieve3000*, a differentiated online program designed to improve reading and writing, met WWC standards with reservations because it was a quasi-experimental design embedded within one district. Another program that met WWC standards was *Fast ForWord Language* (Scientific Learning Corporation, 2004), an interactive computer-based instructional program. Another program reviewed, *Peer Tutoring and Response Groups* (WWC, 2007) is simply a strategy that teachers can use to reinforce previously taught information by having pairs or groups of students work together on a task. None of the programs reviewed were comprehensive, teacher-led, ELD instructional curricula.

In a randomized controlled trial, Vaughn et al. (2017) examined the efficacy of a content-acquisition and reading-comprehension intervention focused on team-based learning during social studies instruction. The study met WWC standards with reservations and reported that EL students in 8th grade increased their content knowledge and content reading comprehension, but not general reading comprehension. Two of the other studies identified in WWC delivered vocabulary interventions. In one study, the intervention was delivered for 15 minutes daily for the entire school year (Lawrence, Capotosto, Branum-Martin, White, & Snow, 2012). In the other study, an intervention was delivered 45 minutes daily for half the school year (Lesaux, Kieffer, Kelley, & Harris, 2014). Design shortcomings, such as unmatched comparisons (Lawrence et al., 2012) and impacts primarily on researcher-developed measures with no corrections for multiple tests (Lesaux et al., 2014), limited generalizability. Our literature review for students learning English as a second language suggests a paucity of evidence-based language curricula for this crucial topic of instruction and growing group of students (Saunders et al., 2013).

Direct instruction spoken English

Engelmann, Johnston, Engelmann, and Silbert (2010) developed *Direct Instruction Spoken English* (*DISE*) to address the needs of students who speak a wide range of language groups in U.S. schools, from Grade 4 to 12, rather than a few particular language groups. Unique from other ELD curricula, *DISE* teaches spoken English to students who may not speak the same language, which avoids homogeneous grouping students by their native language, a practice not often feasible in middle schools with limited resources and staffing. Instead, *DISE* instructors group students according to their proficiency in the English language, enabling teachers to align instruction to the needs of all students in the group (Estrada, 2014).

DISE has two levels: Level 1 with 100 lessons and Level II with 80 lessons. The authors designed the curriculum for teachers to deliver 90 minutes of daily instruction,

either in one block with a brief break in the middle or two 45-minute periods. Although *DISE* authors specified this amount of instructional time to optimally accelerate students' achievement, 90 minutes of daily instruction was not feasible for the ELD teachers participating in this study. Therefore, districts and teachers committed 45–55 minutes daily to teach one *DISE* lesson every two days. The Method section provides a detailed description of instruction, namely pacing and the number of lessons teachers could complete.

Instructional design and delivery

As noted, there is limited evidence that current ELD curricula have effectively developed EL students' oral language skills (Saunders et al., 2013). We hypothesize that authors of most programs devote limited attention to instructional design principles during development (Stein, Stuen, Carnine, & Long, 2001). Many ELD curricula also lack the specific guidance teachers need to explicitly present activities, provide sufficient practice and review, and integrate students' existing knowledge and skills. When selecting an ELD curriculum to evaluate, we considered the iterative development process and the quality of the curriculum design. We chose *DISE* because of the integrated instructional design, carefully organized scope and sequence of skills, explicit instructional activities, and specific guidance for teachers (Engelmann & Carnine, 1991).

Engelmann et al. (2010) designed the scope of *DISE* from an analysis and selection of content. *DISE* first teaches the most elementary skills across categories of language (i.e., speaking and pronunciation, listening, academic and social vocabulary, and syntax), and then proceeds systematically through additions and alterations that create more complicated applications. The work is cumulative so that newly taught content recurs throughout the program. All subskills needed for more complex speaking tasks are taught early and integrated systematically, first bundled as simple tasks and later embedded in progressively more elaborate conversational skills. The instruction also explicitly focuses students' attention on word meanings and language forms, significant concepts in instructed language learning (Council of Chief State School Officers [CCSSO], 2012; Ellis, 2005). The sequential, structured design of *DISE*, carefully worded examples, student choral responding, and teaching scripts make it easy for ELD teachers to teach the lessons and increase implementation fidelity (Engelmann & Carnine, 1991; Stockard, Wood, Coughlin, & Khoury, 2018). We hypothesized, therefore, that the design of *DISE* might render it more effective than other ELD programs for teaching ELs to speak and understand English proficiently.

Purpose of the study

This study addresses the need for randomized field trials of ELD curricula to evaluate their efficacy in teaching English to students identified as ELs, particularly in middle schools. The study's primary aim was to compare the efficacy of *DISE* Level 1 (*DISE* L1) to the English oral language instruction typically provided in middle school classrooms on the development of students' English oral language skills at the end of one and two years of instruction. We specifically targeted classrooms that taught 6th and

7th grade EL students with beginner to early intermediate level English skills, not EL students with more advanced English skills. We assessed the speaking and listening skills typical of students at the beginner to early intermediate level. We also explored distal language skills that beginning-level students would not usually master within a year or two and reading fluency, a far-transfer skill that is not typically taught by oral language teachers and not included in *DISE*. We then tested whether student English language and reading performance depended on their initial skill with the English language.

Based on principles of learning and retention (Carver & Klahr, 2001) and instructional effectiveness research (e.g., Goswami, 2004; Shaywitz, Morris, & Shaywitz, 2008; Stevens, Fanning, Coch, Sanders, & Neville, 2008), we hypothesized that the careful design and explicit-instruction features of *DISE* would be particularly important during ELD instruction and help EL students acquire the skills that contribute to spoken English language proficiency (Spada & Tomita, 2010). The teacher behaviors—explicit demonstrations and specific, corrective feedback—make clear to students the oral language skills they are learning and how to use them correctly. The specific student behavior—a high rate of independent opportunities to practice and review—helps students gain mastery and fluency with newly learned concepts and skills.

The following sections describe the methods, analysis, and results of this randomized controlled experiment, which answer the research questions below:

1. Do 6th- and 7th-grade ELD teachers who use the *DISE* curriculum, compared to business-as-usual ELD instruction, improve English oral language proficiency among students identified as English Learners with limited, beginner to early intermediate English language proficiency across one school year and two school years?
2. Do middle schools with ELD teachers who use *DISE*, compared to instruction-as-usual, improve distal language and reading outcomes? This exploratory question asks about distal language and far-transfer reading skills that extend beyond instruction typical of beginning-level students.
3. Does student initial skill moderate the relationship between condition and outcomes on measures of language performance?
4. Is there a relation between instructional variables, such as the rates of teacher demonstrations or opportunities for independent student practice, and student outcomes?

Within the continuum of research, the present study falls between efficacy and effectiveness; an efficacy trial in real-world settings "focuses on implementation by indigenous providers in school settings" (Smolkowski, Crawford, Seeley, & Rochelle, 2019, p. 197) and relaxes the experimental control (internal validity) to improve generalizability. This study compares middle schools where English-language teachers taught *DISE* L1 for two years to middle schools where teachers taught their usual curriculum with assessments of teachers' instruction and measures of students' English oral language proficiency. As a cluster-randomized trial, the comparison between schools entails the teachers and students in those schools and administrator, teacher, student, and parent decisions about instruction, English skill, enrollment, attendance, and all other activities related to teacher and student outcomes. Random assignment

is intended to minimize the systematic differences between schools on variables other than *DISE* L1 instruction.

Method

Study overview

The study randomly assigned 29 schools in Texas, Washington, and Oregon, within districts, to either implement *DISE* or conduct business as usual (BAU), a waitlist control condition, in their beginner ELD classrooms. We treated classrooms and schools as the same unit, as only one teacher participated in each school and taught all students in their classroom with either *DISE* or BAU. ELD teachers assigned to *DISE* received training from one certified trainer from the National Institute for Direct Instruction (NIFDI). The teachers all received the same training for implementing *DISE* from this trainer. Teachers in the BAU condition taught the ELD curriculum approved by their district. BAU teachers were offered *DISE* and associated professional development at the end of their study participation.

Figure 1 provides an overview of the research design and sample of students, teachers, and schools. Teachers and students participated for two years. In the following description, *year* refers to the year of teacher and student participation unless otherwise specified. We use *grant year* to refer to the year of the grant. We collected measures on all 6^{th}- and 7^{th}-grade students nested within intact beginner ELD classrooms in the fall and spring of Years 1 and 2. We included any students who joined the classrooms over the two years and continued to assess students who moved out of study classrooms if we could locate them within the district.

Participants

Schools

As presented in Figure 1, we randomized 42 schools in 10 districts to condition. We lost one district with 11 schools because a new administrator declined participation after random assignment. One district with one school in each condition had too few students who met study criteria. Of the 29 middle schools in the final sample, 14 received *DISE* and 15 taught BAU. The sample included 25 schools in Texas, two in Washington, and two in Oregon. Schools joined the study in waves, one in each of the four grant years.

Teachers

Overall, 36 teachers participated in the study. All schools had only a single beginning-level ELD teacher. One *DISE* teacher enlisted his instructional assistant to teach the lower skilled beginner students while he taught the higher skilled beginner students; both received *DISE* training and taught with *DISE*. In five schools, a participating ELD teacher left (e.g., for maternity leave) and was replaced by another teacher (sometimes more than once) for at least part of the study. Across the two years of participation in each school, a total of 17 intervention teachers and one instructional assistant

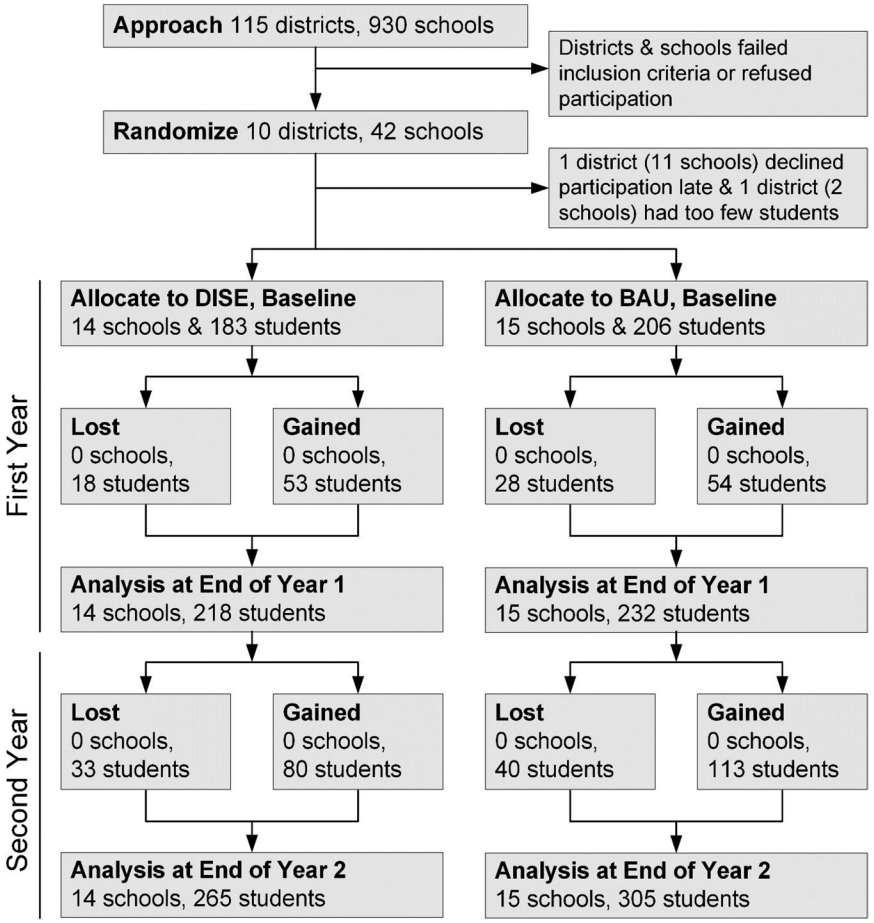

Figure 1. Participant flow diagram.

delivered *DISE* in their ELD classrooms and 19 BAU teachers delivered their usual instruction. All teachers but one, the instructional assistant in a *DISE* school, were certified to teach English to nonnative English speakers (e.g., ESL, TESOL). Teachers taught an average of 11.6 years in total (11.9 in DISE schools, 11.3 in BAU) and 11.8 years of ESL (8.3 in DISE, 14.5 in BAU). In the *DISE* schools, 10 teachers reported that they had received their bachelor's degree, four a master's, and one an associate's. In BAU schools, nine teachers reported a bachelor's, seven a master's, one a doctoral candidacy, one a law degree, and one did not report a degree. One teacher in each condition had a special education certification.

Students

A majority of the 6th- and 7th-grade students spoke Spanish as their first language. See Table 1 for additional information, including sample sizes by year of participation. We assessed 144 students during the fall and 130 students in the spring of Grant Year 1; 366 students in the fall and 417 in the spring of Grant Year 2; 324 students in the

Table 1. Characteristics and measure descriptives for the student sample (N = 746) by condition.

		Control	DISE
Total Sample *N*		412	334
Participation by	Year 1 Fall (T_1) *N* (%)	212 (51%)	185 (55%)
Assessment Period	Year 1 Spring (T_2)	236 (57%)	223 (67%)
	Year 2 Fall (T_3)	298 (72%)	246 (74%)
	Year 2 Spring (T_4)	310 (75%)	265 (79%)
Female *N* (%), 121 missing		171 (48%)	131 (48%)
Ethnicity, 132 missing	Hispanic *N* (%)	290 (83%)	221 (84%)
	Hispanic White	193 (55%)	127 (48%)
	Hispanic AIAN	67 (19%)	81 (31%)
	Non-Hispanic White	14 (4%)	19 (7%)
	Non-Hispanic Asian	25 (7%)	7 (3%)
Free or reduced-price lunch *N* (%), 163 missing		327 (99%)	252 (100%)
Special education *N* (%), 114 missing		9 (2%)	4 (1%)
Grade point average	Year 1 *M* (*SD*) *N*	2.9 (0.6) 194	2.8 (0.8) 179
	Year 2	2.5 (0.6) 181	2.5 (0.6) 161
Days in school *M* (*SD*) *N*	Year 1 *M* (*SD*) *N*	150 (35) 152	143 (38) 129
	Year 2	160 (22) 186	159 (24) 149
IPT *M* (*SD*) *N*	T_1 *M* (*SD*) *N*	8.6 (8.2) 206	7.3 (6.5) 183
	T_2	11.5 (10.7) 232	10.8 (7.6) 218
	T_3	13.6 (13.5) 298	12.2 (8.1) 245
	T_4	15.6 (12.9) 305	14.9 (9.6) 265
ORF WCPM	T_2 *M* (*SD*) *N*	61.1 (32.7) 233	58.1 (27.8) 218
	T_4	74.6 (33.6) 307	73.1 (28.5) 264
WJ Test 1W Scores	T_2 *M* (*SD*) *N*	429.7 (19.8) 230	429.3 (17.7) 217
	T_4	437.4 (21.6) 306	436.5 (18.1) 264
WJ Test 2W Scores	T_2 *M* (*SD*) *N*	433.9 (26.8) 231	429.8 (25.6) 218
	T_4	445.5 (28.7) 306	443.5 (26.5) 265
WJ Test 6W Scores	T_2 *M* (*SD*) *N*	459.9 (22.2) 232	460.7 (19.8) 217
	T_4	466.3 (22.2) 294	469.2 (18.4) 258

Note. Percentages represent the proportion of the total (row 1); students with missing information not included in percentages. Ethnicity reported for only categories with more than 5% of cases in either condition. AIAN = American Indian and Alaska Native. Participating schools were in session for 172 to 177 days. Days in school calculated from days enrolled minus absences. IPT = Idea Proficiency Test; ORF WCPM = oral reading fluency words correct per minute; WJ = Woodcock-Johnson (#1 picture vocabulary, #2 comprehension, # 6 directions.)

fall and 376 in the spring of Grant Year 3; and 95 students in the fall and 107 in the spring of Grant Year 4.

Procedures

Recruitment and Randomization

In the spring and summer before each year of the study, the investigators contacted districts with concentrations of middle school ELs. We focused our recruitment efforts on the West Coast, including Washington, Oregon, California, Arizona, and because of their large concentration of ELs, we also contacted Texas school districts. Middle schools eligible to participate must have (a) used an English language curriculum other than *DISE*, (b) agreed to use their BAU curriculum or teach *DISE* daily for two years, and (c) had at least 10 EL students in grades 6–7 with limited English oral language proficiency. Across the four years of the study, we contacted 115 districts, with 103 districts either not responding to the initial contact or declining after reviewing the materials. We recruited schools in the year before they participated in the project. Ten districts agreed to participate, but two left after random assignment, leaving eight. Schools most often declined due to participation in other research projects, insufficient

number of ELs, new administrators, and concerns about using a new English oral language curriculum. Although all schools had 10 or more ELs when recruited, some had fewer in their first year of participation; one school had only five ELs.

Assessor Training

[Trainer masked] trained research assistants to collect student performance data prior to each assessment period to ensure consistent, standardized administration of the measures. For the initial training, she provided a 4-hour interactive PowerPoint presentation. For returning assessors, [trainer masked] conducted 1.5- to 2-hour refresher trainings. During both initial and refresher trainings, each assessor established reliability at or above 80% agreement with the trainer on each test, and newer assessors verified their reliability through shadow scoring with experienced assessors on the first day of each assessment period.

Classroom observation training. We conducted three observations per year of the ELD instruction provided by DISE and BAU teachers. In Grant Year 1, [trainer masked] trained three observers to use the observation tool, the Classroom Observations of Student–Teacher Interactions (COSTI; Smolkowski & Gunn, 2012). The initial 2-hour training introduced the coding system and codebook, with guided practice coding videos of ELD instruction. Observers practiced coding individually and as a group to compare reliabilities and to adjust decision rules as needed. They then visited a local middle school ESL classroom twice to practice coding in a real classroom and establish final reliability estimates. See the Measures section for a detailed description of the COSTI, the rationale for using it, and the observation timeline.

ELD instruction in study conditions

Teachers delivered classroom instruction primarily in English to all students in the classroom. All teachers taught for their entire class period in middle schools, typically determined at the district level, balancing instructional time and context across intervention conditions. The length of ELD class periods was the same in both conditions, so students received the same opportunity for instruction each day. Teachers taught all students in their classrooms according to condition assignment.

DISE training, coaching, and instruction for intervention schools

High-quality professional development is essential to ensure fidelity of implementation (Fixsen, Naoom, Blase, Friedman, & Wallace, 2005; [citation masked]) and to guarantee that teachers implement curricula well, especially during an efficacy trial (Odom, 2009). In Year 1 of their participation, teachers assigned to the *DISE* condition received two days of training from a trainer certified by the National Institute for Direct Instruction. The trainer provided the standardized training that teachers receive when they purchase *DISE*. The *DISE* teachers learned the instructional skills they needed to teach the exercises as intended, including presentation techniques (e.g., quick pacing, clear demonstrations) and monitoring and correction techniques. In Year 2, the same trainer provided an on-site, half-day refresher training for the teachers, observed them teaching

DISE, and provided up to four online coaching sessions. The coach observed real-time teaching with *DISE* lessons and then debriefed teachers after class and provided feedback. For each online coaching session, the coach provided teachers (and researchers) with specific feedback on the presentation of the lessons, which included lesson pacing, modeling new skills, correcting student errors, and providing adequate practice and feedback for students.

DISE teachers instructed with *DISE* L1, which targets beginner or early intermediate students, and progressed through the lessons as students mastered the content. Students learned how to use English morphology, understand and use basic syntax, pronounce progressively more difficult English sounds, practice complex sentence patterns verbally, and use academic and social vocabulary. All lessons asked students to apply skills in conversation.

DISE instruction. One ELD teacher in each treatment school delivered the *DISE* instruction, which ensured that the intervention condition modeled the actual teaching conditions of typical end-users. The developers of *DISE* intended for teachers to teach 90 minutes daily. However, the teachers in this study had only 45–50 minutes daily to teach their beginners English oral language instead of the recommended 90 minutes. At this pace, we expected them to complete one *DISE* lesson every one to two days. The weekly Lesson Progress Chart (LPC, described below) gave the DISE coach and researchers ongoing data on whether teachers met this goal and barriers teachers faced to providing daily instruction. The number of lessons per week ranged from 0.7 to 1.5 in Year 1 of their participation and from 0.4 to 2.0 in Year 2. On average, students received about 32 lessons, from 19 to 50, in Year 1 and about 37 lessons, from 14 to 69, in Year 2. This level of instruction was less than half of what we expected, on average, in a 33- to 36-week school year. Teachers gave several reasons for not teaching *DISE* daily, including school closure (19%), teacher absences (22%), testing (28%), school events (8%), *DISE* unit tests (4%), and other reasons (19%).

Instruction in BAU classrooms

Across all years of the study, teachers in BAU classrooms reported using the commercially published ELD programs their district had approved for primary English language instruction. Teachers taught with *Milestones*, *Language Power*, *Florida Center for Reading Research Student Activities*, *Keys to Learning*, *Texas Primary Reading Inventory*, and *ESL Reading Smart*. They also used the software programs *Rosetta Stone*, *Imagine Learning*, and *Read 180*. We coded the recordings of our classroom observations to document the instructional content and activities in BAU classrooms to compare with instruction in the DISE classrooms. The Teacher Measures section provides details of the observations and audio content codes. After BAU schools participated for two years, we offered their teachers the *DISE* curriculum and two days of training. Eight BAU teachers received the training and the curriculum.

Teacher measures

Teacher demographic survey

We collected demographic information for all participating teachers and descriptive information on their beginner ELD class during each year of their participation. Demographic questions included education, years of teaching middle school, and years teaching English Learners. ELD class information included curriculum or approach used to teach English oral language, time spent daily on instruction, and use of instructional assistants.

DISE lesson progress chart

We documented the dosage and pacing of *DISE* instruction with a weekly Lesson Progress Chart (LPC) that asked teachers to report the lesson number and the *DISE* activities they completed each day. Teachers also reported the days and reasons why they did not teach *DISE*. A summary of their responses is provided in the Implementation of Instruction Section.

Fidelity of DISE implementation

In addition to the *DISE* LPCs, which tracked the pacing of instruction, we documented fidelity of implementation by comparing the audio recordings of instruction to the lessons as written in the Presentation Books (scripted *DISE* curriculum). We coded each exercise for (a) how well the teacher completed each step of the sequence and (b) adherence to the lesson scripts as written. Overall, teachers followed the lessons and script with fidelity. For more than 90% of the exercises observed, teachers covered all or most of the steps (no more than two skipped or reordered prompts per exercise) and the script (no more than five skipped or reordered words). The *DISE* coach also provided reports of the specific feedback provided to teachers.

BAU teacher survey

We documented the ELD instruction in the BAU classrooms with a quarterly survey that asked teachers to report the ELD curricula they used, their instructional focus, and other considerations that affected instruction (e.g., substitute, testing).

Classroom observations of student–teacher interactions (COSTI). In the fall, winter, and spring of each Grant Year, we documented the rates of specific student–teacher interactions during DISE and BAU instruction. We used the COSTI ([citation masked]; Smolkowski & Gunn, 2012) because it documents the following aspects of classroom instruction not captured by other observation instruments: rates of (a) teacher demonstrations, (b) student independent practice, (c) student errors, and (d) corrective feedback from the teacher. Research and theory suggest that the interactions play a pivotal role in how well students learn and remember basic skills. Observers coded each time that one of the four interactions occurred during 30 minutes of instruction. We recorded audio for subsequent content coding.

Two observers coded instruction for about 20% of observations to establish reliability and maintain reliability at 80% or above. Observers demonstrated a very high

level of interrater reliability; intraclass correlations (ICCs) ranged from .93 to .99. The stability of observed codes over time was lower, with ICCs from .43 to .80. ICCs for stability document the extent to which teachers' rates of instruction practices remain the same over time and are analogous to test-retest correlations. ICCs of 0.21–0.40 describe fair reliability, 0.41–0.60 moderate, and 0.61–0.80 substantial (Donner & Eliasziw, 1987); alternatively, with three observations, an ICC of .40 implies reliability of .67 for an aggregate mean, and an ICC of .50 implies reliability of .75 (Shoukri, Asyali, & Donner, 2004). Consistent with our hypothetical model, we focused on demonstration and independent practice rates, which produced stability ICCs of .57 and .80.

Observers conducted the *COSTI* observations in person during Grant Year 1. In Grant Year 2, we assessed the feasibility and accuracy of remote observations via video conferencing systems with two teachers in the intervention condition and two in the BAU condition. We compared in-person coding to coding online and averaged 80% agreement or above for teacher demonstration and independent practice. We conducted remote observation in the spring Grant Year 2 in all project classrooms. We conducted two of the three annual classroom observations in person in subsequent grant years, on average, and one remotely. The video conferencing platforms enabled accurate *COSTI* coding and also provided audio recordings for later coding.

Audio content coding of classroom observations

Whereas the *COSTI* observations document rates of student-teacher interactions, the *COSTI* was not designed to describe instructional content (focus) and instructional activities. For that reason, we also coded the audio recordings of the observations to describe the aspects of *DISE* and BAU English language instruction not captured with the *COSTI*. We chose the content codes by reviewing the teaching objectives for district-adopted English Learner language curricula and developed a set of codes that described the instructional content and the instructional format (i.e., of typical middle-school English language instruction). For each audio-recorded lesson, we coded 30 one-minute segments of the lesson for the following content areas and activities: (a) phonology; (b) morphology; (c) grammar; (d) syntax; (e) vocabulary; (f) school learning content and abstract concepts; (g) text reading; (h) other low-frequency content, such as discussions, literary forms, or reading aloud; (i) general directions and transitions; (j) teacher-directed activities; (k) peer to peer activities without teacher facilitation; (l) teacher-directed conversation; (m) independent seatwork; (n) teacher-directed writing; (o) teacher or student talk in Spanish; and (p) computer or audio assisted instruction. For most codes, however, we found insufficient variability for analysis. In addition, several were included to capture time in class but were deemed unrelated to *DISE* instruction (e.g., seatwork, writing, computer instruction). We examined only vocabulary, syntax, and teacher-directed activity. Coders demonstrated a very high level of interrater reliability; ICCs ranged from .88 to .92. The stability of observed codes over time was lower, with ICCs of .15, .58, and .56 for vocabulary, syntax, and teacher-directed activity, respectively. Table A1 in the appendix reports the codes by condition with frequencies higher than 2%.

Student measures

Considering school district concerns for testing time, we opted to focus on the following battery of measures. We selected the IDEA Proficiency Test (IPT) as our primary measure of oral English language. We also administered three Woodcock-Johnson (WJ) language measures and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS 6th ed.) measure of oral reading fluency (ORF) as experimental measures of distal outcomes. We did not expect changes in ORF, for example, but included it to assess the potential for student oral language skills to transfer to reading skills. The WJ language measures are more proximal to *DISE* instruction, as they measure English language, but assess concepts and skills that beginning-level students would not likely master within a year or two. Therefore, we focused our investigation on the IPT and considered the WJ and ORF as exploratory measures because they assess skills likely beyond the level of *DISE* instruction provided to students in this project.

English oral language proficiency

We established students' baseline English oral language proficiency using the IPT assessment suite (Ballard & Tighe, 2010a). We administered the IPT II English Oral Language Test at the beginning and end of the school year. The IPT II English Oral Language Test is an individually administered assessment that takes approximately 5 min for students with lower English language proficiency and 40 min for students with intermediate to early advanced English proficiency levels to complete (Ballard & Tighe, 2010a). IPT II English Oral Language has two alternate forms; we used one in the fall and one in the spring (Ballard & Tighe, 2010b; 2010c). According to the publisher, the internal consistency reliability of both forms of IPT II Oral English, E and F, is .90 (Ballard & Tighe, 2010a). The test–retest reliability was .95 (Ballard & Tighe, 2010a). Because the IPT was our primary outcome measure, we collected detailed reliability data on 161 occasions. One assessor administered the IPT to a student and the other assessor shadow-scored to measure reliability. We estimated ICCs greater than .99 for items correct and .95 for items incorrect, which suggests "almost perfect" (Donner & Eliasziw, 1987, p. 443) interrater reliability.

We also administered the Woodcock-Johnson IV Tests of Oral Language (*WJ IV;* Schrank, Mather, & McGrew, 2014), a well-established, reliable, and valid measure of oral language. The Oral Language cluster measures expressive and receptive language, vocabulary, listening comprehension, linguistic competency, reasoning, and memory. We used the picture vocabulary, oral comprehension, and understanding directions subtests as a distal measure of student growth in vocabulary and listening comprehension for this study.

English oral reading fluency

Oral reading fluency (ORF) is a valid, reliable indicator of overall reading competence (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Slocum, Street, & Gilberts, 1995; Stanovich, 2000). We assessed students' ORF rate at spring data collection with one 6th grade passage (DIBELS 6th ed.; Good & Kaminski, 2002). We calculated the number of words correctly read in 1 min. DIBELS ORF has been shown to predict later reading

proficiency (e.g., Baker et al., 2008; Good & Kaminski, 2002). [Citation masked] (2016) and Cummings, Smolkowski, and Baker (2021) demonstrated the accuracy of DIBELS 6th edition ORF when predicting comprehensive reading tests.

District measures

We also collected student data from districts, including grades, attendance, special education status, and free or reduced-price lunch rates (details included in the appendix).

Statistical analysis

To test for condition differences, we compared *DISE* and BAU schools within multilevel models that assessed net gains across the first school year or growth across two school years (Fitzmaurice, Laird, & Ware, 2004; Murray, 1998). The analyses of gains and growth avoid assumptions required by approaches such as ANCOVA, which may be untenable (Allison, 1990; Willett, 1988). Gains and growth models incorporate baseline information into the intercept (Fitzmaurice et al., 2004; Murray, 1998; Zvoch & Stevens, 2006). Growth models do not require fixed spacing of assessments and allow piecewise growth (i.e., splines; Fitzmaurice et al., 2004) and flexible specification of the covariance among repeated assessments (e.g., autocorrelation; Wallace & Green, 2002). Growth models also accommodate students missing a subset of assessments, "a major advantage" (Hox, 2010, p. 80) because it reduces bias due to missingness.

Efficacy across the first year

The IPT represented the primary outcome and the only measure collected at all four time points. The model to test the efficacy of *DISE* during the first year included fixed effects for Time, Condition, and the Time × Condition interaction, which tests net gains over the first year (from T_1 to T_2; Murray, 1998). We coded Condition 1 for *DISE* schools and 0 for control schools; Time was coded 0 at T_1 and 1 at T_2. With this coding, the Time × Condition effect estimates the difference in gains between the two intervention conditions. The models nest the student parameters within schools to account for clustering effects (Murray, 1998).

Efficacy across two years

Growth models for two years are more complicated than those for one year. *P* values summarize the probability of a test parameter correctly only if the model assumptions are correct (Greenland et al., 2016). With fall and spring assessments across two years, linear growth is unlikely to model the underlying data correctly. We therefore tested condition differences in growth across two years (T_1 to T_4) with an approach described in Low, Smolkowski, Cook, and Desfosses (2019). We defined models that varied in their fixed effects and within-student covariance structures, including linear growth. For example, one model tested within-school-year growth that allowed for improvement only during the school year but not during the summer. We then fit the data to each model and selected the best fitting model before interpreting intervention effects.

We described the model-selection approach in more detail within the appendix. The best-fitting model assumed linear growth for students in the BAU condition but within-school-year growth for students in the *DISE* condition. This model allowed students in *DISE* schools to differ from students in BAU schools during the school year, between T_1 and T_2 and between T_3 and T_4, but assumed equivalent gains during summer. After selecting the best-fitting model, we then inspected and interpreted condition effects.

Moderation

DISE targeted students with very limited English. We expected that initial language skills would moderate condition differences. Specifically, we did not expect students who began the study with intermediate or higher English language skills to benefit from the beginner-level *DISE* L1 instruction geared toward beginner students. To test pretest IPT scores as a moderator, we added baseline IPT and its interaction with the Time and Time × Condition terms in the first-year model and the selected growth model for two-year outcomes.

Exploratory analyses

We explored differences between conditions on four additional measures: ORF and, from the WJ Oral Language Cluster, picture vocabulary, oral comprehension, and understanding directions. Because these measures were collected at T_2 and T_4 only, we could not analyze gains or growth, so we analyzed these data with a mixed-model ANCOVA using baseline IPT scores as a covariate. We explored moderation in the ANCOVA model by adding a pretest IPT × Condition term.

Model estimation and missing data

We fit the statistical models to our data using SAS PROC MIXED version 14.2 (SAS Institute, 2017) with full-information maximum likelihood estimation (FIML). For exploratory measures, we fit ANCOVA models with multiply imputed data. Consistent with the design and intent of the study, we conducted a cluster-focused, intent-to-treat analysis (Vuchinich, Flay, Aber, & Bickman, 2012) in which the primary analysis incorporated all available data. This approach includes late entrants (joiners) to increase generalizability and reduce the potential for bias (Brown et al., 2008; Vuchinich et al., 2012).

The gains and growth models that tested condition differences with the IPT included all cases with data at any time point (Allison, 2012; Graham, 2009) and relied on FIML estimation. The ANCOVA models that tested exploratory variables relied on multiple imputation using Markov chain Monte Carlo methods and the expectation-maximization algorithm with 100 imputations (Graham, Olchowski, & Gilreath, 2007) in SAS PROC MI. ANCOVA models are not suitable for FIML estimation because analysis software deletes all cases with missing covariates. Multiple imputation included all variables used in the analyses and student attendance, enrollment, and grade point average from both school years. The approach did not include an intervention condition indicator in the model or impute separately by condition.

Although WWC (2020) recommended accounting for intervention condition, it may spuriously inflate condition differences (Smolkowski, Danaher, Seeley, Kosty, & Severson, 2010). We employed Graham's (2012) hybrid dummy-code strategy to account for clustering, with fixed effects for each school (except one) in 25 of the 100 imputations. The remaining 75 imputed data sets did not include the dummy codes.

We conducted multiple sensitivity analyses for our missing data approach. The first varied the variables included during the imputation process. The second sensitivity analysis varied the proportion of imputations that accounted for clustering (10% to 50%). We also analyzed complete cases, which entails the listwise deletion of all students without complete data. Although the WWC (2020) recommends complete-case analysis, the approach is more, not less, likely to introduce bias. A complete case analysis can produce biased estimates with a large proportion of missing cases, as in the present sample, and fails to use all available information (Allison, 2009; Graham, 2009). None of the sensitivity analyses changed the interpretation of the results. We describe our missing-data approach further in the appendix.

Reporting

In response to the recommendations of the American Statistical Association (Wasserstein, Schirm, & Lazar, 2019; Wasserstein & Lazar, 2016), we abstained from using hard cutoffs and claims of "statistical significance" when $p < .05$ or for other metrics (e.g., effect sizes). We reported p values, Benjamini-Hochberg adjusted p values (p_{BH}), and Hedges's g values along with their 95% confidence intervals (CI) to describe the results (for dichotomous variables, we calculated Cox's d instead of Hedges's g). To supplement these standard statistics, we estimated model probabilities to characterize the strength of evidence for the alternative hypothesis over the null hypothesis. P values have a cumbersome definition as a measure of incompatibility between the observed data and all statistical model assumptions, including the null hypothesis, H_0 (Wasserstein & Lazar, 2016). They inform on neither which assumptions are incorrect nor the importance of the association (Greenland et al., 2016). The model probability, w, describes the strength of evidence for the hypothesis of an intervention effect (Burnham, Anderson, & Huyvaert, 2011). We defined a model for each of two hypotheses, the hypothesis of an intervention effect (H_A) and the hypothesis of no intervention effect (H_0), and reported the model probability for the model with the condition effect (H_A). With only two models, the model probability for H_0 (no condition effect) is $1 - w$. For example, if $w = .75$, it suggests the probability of H_A is .75 while the probability of H_0 is .25. Equivalently, the model for H_A is only three times as likely as the model for H_0. We do not use a particular probability level as a cutoff for "significant" but interpret them as a continuous indicator of evidence for intervention effects.

Results

Descriptive results and baseline equivalence

We report demographic characteristics in Table 1. Students in the *DISE* and BAU conditions did not meaningfully differ in proportions of female students (Cox's $d = 0.01$)

or the proportion of Hispanic students ($d = 0.04$). We found larger differences between conditions for the proportion of students who were Hispanic White ($d = 0.17$) and Hispanic American Indian Alaska Native ($d = 0.38$) but did not examine other racial or ethnic categories due to small samples, less than 5% in one or both conditions. The proportion of students in special education also differed by condition ($d = 0.32$). Cox's d is very sensitive, however, to differences between groups low base rates (< .10 or > .90). The 0.32 standard deviation difference for special education status represents nine students in the BAU condition and four in the *DISE* condition, numbers too small to affect results meaningfully. Because 100% of students in the *DISE* condition qualified for free or reduced-price lunch, we could not calculate d.

Table 1 also provides descriptive statistics for the IPT and our four experimental outcomes for all assessment times. Students in BAU schools scored more highly on the IPT than students in *DISE* schools at pretest, Hedges's $g = 0.17$. Notably, the sample included students who scored higher than expected on the IPT. This study, and *DISE* L1, targets beginner to early intermediate students likely labeled Level A or B on the IPT. These levels represent students below the 20th normative percentile with IPT scores of 14 or below. In our sample, 25% of the students scored 12 or above in the fall of Year 1 or 18 or above in the fall of Year 2. At least one student scored at the 69th percentile, 43, in the fall of Year 1 and at the 96th percentile, 76, in the fall of Year 2. Although we hypothesized that baseline IPT might moderate condition effects, the higher-than-expected scores punctuated the underlying research question.

Efficacy across the first year

We tested the efficacy of *DISE* from the fall to spring in each school's first year of implementation. The analysis of *DISE* impact began with an examination of attrition and joiners. We then tested for condition differences.

Attrition and joiners

The analysis of joiners and attrition included inspection of the proportion of students missing in the first year at baseline or the Year 1 posttest (T_2). We also tested for differential *effects* of joiners or attrition with a mixed-model analysis of variance. Specifically, we regressed T_2 or T_1 IPT scores on (a) study condition, (b) missingness status, and (c) the interaction between the two (Graham & Donaldson, 1993). We summarize the results next and provide additional detail in the appendix.

Joiners (late entrants) represented approximately 24% of the Year 1 sample, which differed between conditions by less than 3%. The multilevel test of differential effects indicated that joiner effects were unlikely (IPT interaction = −0.53, 95% CI [−4.31, 3.26], $t_{27} = -0.29$, $p = .7769$, $w = .27$). Here, w represents the probability of the hypothesis that includes the missing-by-condition interaction compared to a hypothesis without the interaction, which suggests that the hypothesis of differential effects due to late entrants was unlikely. At T_2, the IPT difference between conditions for students without pretest data, $g = -0.17$, was larger than for students with baseline data, $g = -0.03$. Attrition at T_2 represented 12% of students present at T_1, with a net difference between conditions of 4%. The test of differential effects of attrition by condition was equivocal

(IPT interaction = 3.58 [−1.01, 8.17], t_{27} = 1.60, p = .1214, w = .56). Baseline IPT differences between conditions were $g = -0.65$ for students missing T_2 data and $g = -0.11$ for students with data at T_2.

Given that relatively few students were missing data, interpretation of effect sizes for that sample requires caution. They are likely less stable than estimates from the larger group of students with data. The effect sizes were also larger for students missing data, at either time point, than those with complete data, suggesting that, to some extent, joiners replaced similar dropouts with respect to IPT scores. Full-information maximum likelihood with all available data and multiple imputation, used for analyses discussed below, help minimize the potential for bias, likely more so than complete case analyses (Allison, 2009; Schafer & Graham, 2002).

Main effects

We anticipated *DISE* would lead to improved IPT scores for students in the intervention condition, but the Time × Condition model did not support that hypothesis: IPT difference = 0.51, 95% CI [−1.13, 2.16], $g = 0.05$ [−0.12, 0.23], t_{27} = 0.64, p = .5282, w = .30 (see Table 2). The model probability suggests that the data best fit the model without condition differences.

We conducted a complete case analysis to test the sensitivity of the results to missingness assumptions. These analyses confirmed the results that relied on all available data: IPT difference = 0.52 [-0.92, 1.97], $g = 0.06$ [−0.10, 0.21], p = .4625, w = .32 (see Table A2 in the appendix for all full model results). We again advise caution when interpreting results from a complete case analysis (Allison, 2009; Graham, 2009). In the present case, the standard errors and confidence intervals from the complete case analysis are smaller, suggesting that the analysis with only those students who contributed data at both time points may have underestimated the variability in IPT scores when compared to the full sample.

We explored condition differences on four additional measures with mixed-model ANCOVAs: ORF, picture vocabulary (WJ Test 1), oral comprehension (WJ Test 2), and understanding directions (WJ Test 6). None of the exploratory measures demonstrated meaningful differences between conditions (Table 3; see the appendix for additional details).

Moderation

We anticipated that students who began the year with higher performance on the IPT would not benefit as much from *DISE* as students who began the year with lower levels of English language proficiency. We therefore tested for differential response on the IPT at T_2 to *DISE* (moderation) based on baseline IPT scores. Students scored similarly across conditions on the IPT at baseline (see Table 1) and most notably at the extremes, where analyses become sensitive to outliers. For example, about 5% of students in *DISE* schools and 6% in control schools scored a zero on the IPT. Across conditions, about 90% of students scored between 1 and 19. The 95th percentile and maximum values were 18 and 41 in *DISE* schools and 20 and 43 in comparison schools. Hence, the distribution of baseline scores or extreme outliers did not likely unduly influence the moderation tests.

Table 2. Impact of DISE on IPT from a time × condition models for year 1, from T_1 to T_2, and for growth across year 1 and year 2, from T_1 to T_4, with all available data.

Statistic or parameter estimate		IPT Gains from T_1 to T_2	IPT Growth from T_1 to T_4
Model probabilities		.31	.86
Fixed effects	Intercept	8.56 (1.45)	7.85 (1.34)
	Time	4.58 (0.56)	3.08 (0.33)
	Condition	−1.70 (2.09)	−1.64 (1.93)
	Time × Condition	0.51 (0.80)	1.27 (0.52)
Variances	School-Level Intercept	25.43 (8.89)	23.58 (7.58)
	School-level Gains	.92 (0.68)	1.91 (0.72)
	Student-Level Intercept	43.84 (3.58)	
	Residual	14.30 (1.15)	
	T_1		88.71 (5.24)
	T_2		87.09 (5.12)
	T_3		72.69 (4.60)
	T_4		43.00 (2.90)
Hedges's g†	Time × Condition	0.05 [-0.12, 0.23]	0.14 [0.02, 0.25]
P value	Time × Condition	.5282	.0222
Degrees of freedom		27	27

Note. Model estimates relied on all available data. Cells with parameter estimates contain the estimate and, in parentheses, standard error. Time coded 0 at baseline, 1 at T_2, 2 at T_2, and 3 at T_4. Condition coded 1 for schools that taught DISE and 0 for control schools. For the IPT gains model from T_1 to T_2, the student-level intercept can also be interpreted as the covariance between T_1 and T_2, and the residual can also be interpreted the variance of the gains (see Murray, 1998). In the IPT growth model from T_1 to T_4, the Time × Condition term assumed no growth for the DISE condition during the summer, between T_2 and T_3, over and above normative growth estimated by Time for the control sample; Time × Condition was coded from T_1 to T_4 as 0, 1, 1, 2. The growth model included independent variance estimates for each time point, as well as covariances between assessments (not shown). Although the data fit the unstructured model best, when compared to a compound symmetric or autoregressive variances, the fixed effect for Time × Condition differed only slightly between models.

† Hedges's g for IPT Growth from T_1 to T_4 represents the average per-year effect of DISE (g = .28 for two years).

The model produced a moderation effect of −0.27 [−0.43, −0.12], t_{27}= −3.58, p = .0013, w = .99. The model probability, w, compared the models with and without the moderation effect and supported moderation. The model results, displayed in Figure 2 (top), show that for students with an IPT score below 5.5, those in *DISE* schools scored higher on the IPT at T_2 than those in BAU schools. This suggests an intervention effect for the lower-scoring half of the students. Baseline IPT scores did not produce a differential response to *DISE* for exploratory measures.

Efficacy across two years

Students potentially received *DISE* for two years. We tested for differences between conditions at T_4, the end of Year 2, although the two-year sample included considerably more students missing baseline and T_4 scores. We first report the results of an analysis of joiners and attrition and then the results of the comparison between intervention conditions.

Attrition and joiners

The analysis of attrition and joiners proceeded as for Year 1 but considered joiners by T_4 and attrition for those students present at T_1 but not at T_4. Among joiners by T_4 data, 53% did not have pretest data, but the rates were similar across conditions (5% difference). The test for the condition-by-missingness interaction suggests that a

Table 3. Impact of DISE on exploratory reading and language measures at T_2 and T_4 from ANCOVA with baseline IPT as covariate and the complete-case sample.

		T_2				T_4			
		Oral reading fluency	Picture vocabulary, WJ 1	Oral comprehension, WJ 2	Understand directions, WJ 6	Oral reading fluency	Picture vocabulary, WJ 1	Oral comprehension, WJ 2	Understand directions, WJ 6
Model probabilities		.26	.33	.27	.74	.27	.30	.26	.44
Fixed effects	Intercept	45.90	418.78	416.44	450.75	64.59	429.57	442.63	466.92
		(3.12)	(1.33)	(1.83)	(1.37)	(3.40)	(1.66)	(2.03)	(1.42)
	Condition	−0.52	1.38	−0.59	3.55	0.95	1.29	0.24	2.10
		(3.85)	(1.58)	(2.18)	(1.61)	(4.08)	(2.04)	(2.31)	(1.64)
	Baseline IPT	2.34	1.77	2.64	1.72	2.07	1.70	1.75	1.13
		(.18)	(.09)	(.12)	(.09)	(.21)	(.10)	(.14)	(.10)
Variances	School Level	49.74	5.08	9.23	5.10	33.82	11.35	2.51	2.55
		(26.69)	(4.41)	(9.53)	(5.00)	(27.32)	(7.62)	(9.84)	(4.95)
	Residual	537.46	131.49	257.48	140.07	668.93	133.50	319.47	136.69
		(43.14)	(10.60)	(20.91)	(11.33)	(59.98)	(12.12)	(28.95)	(12.71)
ICC		.09	.04	.04	.04	.05	.08	.01	.02
Hedges's g	Condition	−0.02	0.08	−0.02	0.21	0.03	0.07	0.01	0.14
P value	Condition	.8935	.3901	.7871	.0364	.8177	.5333	.9196	.2127
P_{BH} value	Condition	.8935	.7802	.8935	.1456	.9196	.9196	.9196	.8508
Degrees of freedom		27	27	27	27	24	24	24	24

Note. Model estimates from complete-case sample, with 332 to 336 students and 27 degrees of freedom (*df*) for T_2 measures and 257 to 270 students and 24 *df* for T_4 measures. Condition coded 1 for schools that taught DISE and 0 for control schools. The analyses with multiple imputation produced results that were similar but often less supportive of condition effects and could not produce model probabilities. Cells with parameter estimates contain the estimate and, in parentheses, standard error. P_{BH} values are Benjamini-Hochberg corrected *p* values. WJ 1 is the Woodcock-Johnson Test 1 and similarly for WJ 2 and WJ 6. The analyses used W scores for all Woodcock-Johnson measures and correct words per minute for oral reading fluency.

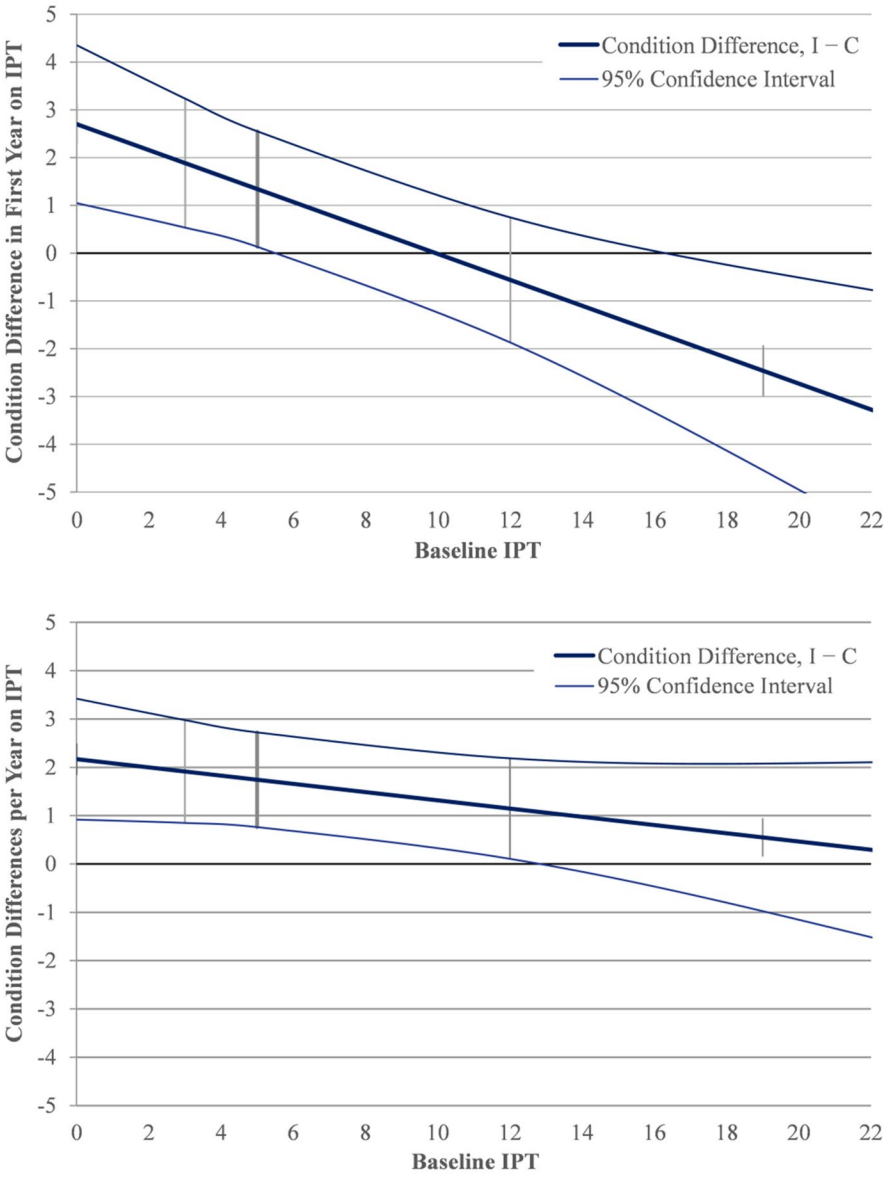

Figure 2. Differential effects of DISE on English language skill (T_2 IPT) over first school year (top) and two years (bottom) based on baseline English language skill (T_1 IPT).

Note. The vertical axis represents the difference between conditions; zero represents no difference. The horizontal axis represents the range of baseline scores and was truncated at 22. The heavy decreasing line represents the estimated mean difference between conditions at each baseline value. The two thinner, outer lines depicts the 95% confidence intervals (CIs) around the mean difference. The vertical lines indicate sample percentiles: the heavy, longer line is the median; the thin, longer lines depict the 25th and 75th percentiles; and the short lines show the 5th and 95th percentiles. The 5th percentile falls at zero on both graphs (not shown) and the median falls at 5.0. In the first year (top graph), the CIs exclude zero for baseline IPT scores below 5.5 (52% of the sample). Across two years (bottom graph), the CIs exclude zero for baseline IPT scores below 12.7 (77% of the sample).

differential joiner effects were unlikely (IPT interaction = 0.85, 95% CI [−2.44, 4.14], t_{27} = 0.53, p = .6016, w = .29). At T_4, students without baseline data differed between conditions, $g = -0.15$, more than students with baseline data, $g = -0.02$, both favoring the BAU condition. In the attrition analysis, we found that 31% of students with data at T_1 did not have data at T_4, and the rate differed by 5% between conditions. An analysis of differential attrition effects suggested that a condition-by-missingness interaction was unlikely (IPT interaction = −0.07 [−3.32, 3.19], t_{27} = −0.04, p = .9660, w = .26). The differences between conditions were slightly smaller for students missing T_4 data, $g = -0.12$, than for $g = -0.19$ for students with data at T_4. In terms of their differences between conditions, students missing at baseline and T_4 largely replaced each other, $g = -0.12$ and −0.15, respectively. See the appendix for additional details.

Main effects

For tests of efficacy on the IPT across two years, we fit a growth model that assumed linear growth for students in BAU schools but a difference in growth between conditions only during the school years, from T_1 to T_2 and from T_3 to T_4. The model anticipated that students in both conditions improved their English language at the same rate during the summer. The results suggested a small probability that students in *DISE* schools made the same progress as those in BAU schools across two years: IPT difference = 1.27 [0.20, 2.34], $g = 0.14$ [0.05, 0.25], t_{27} = 2.43, p = .0222, w = .86. Compared to the model representing the null hypothesis, H_0, the model that included differences between conditions during the summer, H_A, has a probability of .86. Table 2 presents the model estimates for differences in IPT growth over two years. Importantly, the IPT difference in growth, 1.27, and the effect size, $g = 0.14$, represent the impact of *DISE* on language *each year*. Students who participated in both years should expect a quarter-standard-deviation improvement in English language skills ($g = 0.28$).

With considerable missing data, we conducted a complete case analysis to test the sensitivity of the results to missingness assumptions, which confirmed the results with all available data: IPT difference = 1.67 [0.64, 2.71], $g = 0.19$ [0.07, 0.30], t_{27} = 3.31, p = .0027, w = .98. The effect size represents the impact for one year, so the effect size for students who participated in both years would be 0.38 (see Table A2 in the appendix). The effect size from the complete case analysis exceeds that from the analysis of all available data. For students who remained in participating schools for both years and provided data at all assessments, DISE may have provided greater benefit. The larger effect size may also have resulted from reduced sample variability among students with complete data. For these and other reasons (Allison, 2009; Graham, 2009), we interpret the results from the analysis with all available data.

As with tests of intervention effects across only Year 1, we found no evidence that *DISE* improved students' ORF, picture vocabulary (WJ Test 1), oral comprehension (WJ Test 2), or understanding directions (WJ Test 6). See Table 3 for model results.

Moderation

We tested for differential response to *DISE* over time based on students' baseline IPT scores, which were distributed similarly across conditions. The model produced a moderation effect of −0.09 [−0.19, 0.02], t_{27} = −1.63, p = .1146, w = .54. The model

probability, *w*, which compared the models with and without the moderation effect, and the *p*-value offered minimal support for moderation. Because pretest moderated the effects of *DISE* during the first year, however, we depicted the effect for two years in Figure 2 (bottom). The figure shows that for 77% of the student sample, those with initial IPT scores below 12.7, the *DISE* condition generated higher IPT scores per year than those in the BAU condition.

DISE instruction and IPT gains

To examine further whether *DISE* instruction was associated with improved language outcomes, we explored correlations between IPT gains and measures of *DISE* instruction. We estimated standardized regression weights, denoted *r*, from multilevel models that nested students within schools that taught *DISE*. We estimated associations for the number of lessons completed per week by *DISE* teachers and the observed rate of teacher demonstrations, rate of independent practice opportunities, and the proportion of time spent on vocabulary. We chose these measures because they were unique to the *DISE* curriculum and had sufficient variability and reliability. For two-year gains on the IPT, we averaged the predictors across both school years. Given the efficacy tests suggested moderation by baseline IPT, we examined correlations for all intervention students and then separately for those who began at IPT Level A.

We found modest correlations between IPT gains and most predictors. The proportion of time spent on vocabulary predicted IPT gains for both samples and across one and two years (r = .24 to .36). For students who began the study with limited language skills, the rate of teacher demonstrations correlated with gains across the first school year (r = .22). The number of lessons per week predicted IPT gains across two years, T_1 to T_4 (r = .25 to .28). We presented all correlations in the supplemental appendix, Table A3.

Discussion

The primary aim of this study was to compare instruction with *DISE* to the ELD instruction typically provided in 6th- and 7th-grade classrooms on the development of students' English oral language skills at the end of one and two years of instruction. We explored differences in distal language and reading outcomes, and differential response to *DISE* by initial skill on English oral language skills. We also assessed the association between student language performance and number of *DISE* lessons taught per week, rates of teacher demonstrations, independent practice opportunities, and the proportion of time spent teaching vocabulary. We summarize these results, present implications, describe study limitations, and conclude with the significance and generalizability of the results in light of these factors.

Results summary

We hypothesized that *DISE* would more effectively develop beginning English oral language skills compared to typical ELD instruction. *DISE* provides specific instructional

guidelines for teachers, controlled introduction of English language skills, cumulative review, and mastery-based approach, characteristics often absent from standard ELD approaches. We examined the IPT as the primary measure and explored three generalized language measures from the WJ Oral Language Cluster and ORF as distal measures. Overall, the analyses suggest that differences between intervention conditions on English oral language favored the *DISE* condition across two years. In this study, schools that participated in both years exhibited a quarter-standard-deviation improvement in English language skills among their students ($g = 0.28$). This effect size is a small but meaningful difference, given the importance of teaching middle school newcomers English as quickly as possible to access the language used in content area instruction.

As described earlier, we analyzed schools for two years with the students enrolled in those classrooms. Students may have enrolled in the ELD class for both years but possibly in only the first or the second year. The two-year analysis included all of those students, which is important for assessing schools' efforts to improve ELD instruction. Two studies of normative effect sizes can also help contextualize the effects for *DISE*. Hill, Bloom, Black, and Lipsey (2008) reported average effect sizes of 0.24 for reading tests and 0.07 for mathematics across a full year of middle school. Scammacca, Roberts, Vaughn, and Stuebing (2015) more recently found that effect sizes typically ranged from .18 to .30 for middle school students receiving reading interventions. Hence, our gain of 0.28 standard deviations falls in the range of outcomes considered effective.

The analysis of differential response to instruction based on initial English language skills demonstrated that not all students benefited from *DISE*. Students who scored in the intermediate and advanced language proficiency range did not benefit from *DISE* L1 instruction, as we expected, because *DISE* L1 targets beginner proficiency levels. Students who began with lower English language proficiency benefited the most from *DISE* L1 instruction over one and two years, as shown in Figure 2. *DISE* L1, therefore, appeared to benefit most the population of students for whom the instruction specifically targeted.

Among *DISE* classrooms, we anticipated that the number of *DISE* lessons completed per week, the rate of teacher demonstrations, the rate of independent practice opportunities, and the proportion of time spent on vocabulary would predict student outcomes on the IPT. We found limited correlations between these predictors and IPT gains. We also found modest but positive correlations with the proportion of time spent teaching vocabulary and the rate of teacher demonstrations for students who began with limited language skills with student IPT gains. These predictors represent only limited features of intervention intensity (Warren, Fey, & Yoder, 2007) and content, and correlations cannot establish cause or the direction of a relationship. Nonetheless, they are consistent with the conclusion that the differences between intervention conditions were associated with *DISE* instruction.

Study implications

Our hypotheses received modest but encouraging support. *DISE* instruction had its most significant impact on students who initially demonstrated beginner levels of

English fluency, as suggested by the moderation effects (Figure 2). In contrast, students with higher levels of English skills benefited less, if at all. We found this unsurprising given the high fall scores by some students. The *DISE* instruction evaluated in this study targeted students with beginner to early intermediate level of English skills, yet some classrooms included students with much higher English oral language skills than anticipated. Roughly a quarter of the sample scored higher than the expected range for beginner to early intermediate oral language, and some scored above the 50th normative percentile. Anecdotally, we learned that some high-scoring students had received one or more years of English instruction in their home country before arriving in the United States. For students with intermediate to advanced English language skills, placement at a higher lesson in *DISE* Level 1 or Level 2 may have been more appropriate. Districts and schools could not change their placement procedures to ensure ELD classes contained only beginners or early intermediate students. Existing student placement structures dictated that the intermediate to advanced students enrolled in the same classes as the beginners and early intermediate students.

It is important to recognize that investigations of curricula entail the delivery of instruction as well as the specific classroom and school context, which includes the teacher as the agent of delivery, the students as recipients of the instruction delivered by their teacher, and all student interactions during the school year (Raths, 1967). The comparisons between conditions in this cluster-randomized trial necessarily encompass the teachers and students with all their skills and histories. Randomization at the school level aims to balance all factors unrelated to the intervention across conditions; differences between conditions capture all systematically varied features. Systematic differences may have included the *DISE* curriculum, the training and coaching teachers received, differences in instructional delivery, and changes in the students' interactions that resulted from the curriculum and teacher behavior. The differences between conditions, then, are associated with these school and classroom contexts. Interpreting these aggregate results at the individual level is called the *ecological fallacy* (Hox, 2010). It is important to remember that inferences apply to classrooms of students as a whole and not necessarily to individual students.

In this study, intervention teachers experienced several constraints on their teaching time that reduced the number of *DISE* lessons they could teach and the amount of instructional time they could devote to English oral language development. Teachers had 45-55 minutes daily to teach their EL students English oral language instead of the 90 minutes the *DISE* developers recommended. Based on coaching reports, *COSTI* observations, and our comparison of the *COSTI* audio recordings to the lessons as written in the *DISE* presentation books, teachers taught the program with a high level of fidelity when they were able to teach. However, they were rarely able to teach 5 days a week, even for 45-55 minutes. The limits on time represented an ongoing implementation challenge throughout the study that restricted students' exposure to *DISE* and to their opportunities to learn English. Although we expected teachers to complete 2.5 lessons per week, on average, or about 80 lessons during the school year, no teachers reached this goal in either year. The average student received less than half: 32 in Year 1 and 37 in Year 2. Nonetheless, the number of lessons taught per week and the time spent teaching vocabulary predicted IPT gains across two years of

instruction, and students appeared to benefit more from two years of instruction than one. These results suggest that an increased duration of instruction may benefit students.

Limitations

The assessment battery focused on the IPT as the proximal measure and the Woodcock and ORF as more distal measures of English language development. Future research with similar samples would benefit from using more proximal assessments to the intervention content. Not to be confused with instructional fidelity, implementation dosage, such as teaching time and lesson coverage, was an ongoing consideration. Given substantial research support on the effectiveness of Direct Instruction (DI) programs such as *DISE* (e.g., Stockard et al., 2018), students' growth in English oral language in this study was less than expected. The limited effects were likely due, in part, to insufficient instructional time and competing demands on teachers' time.

Heterogeneous grouping of students in *DISE* classrooms also affected implementation. Schools grouped all students considered, loosely, beginners or newcomers, which mixed students with no English skills and those who demonstrated intermediate and more advanced English skills. Developers designed *DISE* to teach ELs with similar English skills, but homogeneous grouping for most schools may be neither feasible nor practical. Most schools had one ELD period per day. Teachers presented *DISE* lessons to their whole class, with a range of English oral language performance levels, thereby limiting their ability to tailor instruction to the needs of individual students. Teachers gave the *DISE* L1 placement test to determine the level of instruction that addressed the instructional needs of most of their students. The placement test asked students to follow basic directions first, such as "stand up" and "touch your head," and progressively presented more challenging language prompts. Teachers discontinued testing when students missed five consecutive items, and the total number correct was used to determine placement (i.e., the initial lesson) in the program. The coach trained teachers to monitor their students' understanding and provide more practice and support for struggling students. However, heterogeneous grouping of students may have slowed down language learning for some students whose skills were higher or lower than the majority of the students in the class. Other strategies, such as small group instruction with paraprofessionals, may allow for additional differentiation.

Conclusion

In order to promote the overarching goal of bilingualism and the quality of life benefits that come with being bilingual and bicultural, schools must have effective English language curricula and training for teachers. The present study suggests that combining an evidence-based English oral language program and daily instruction can produce meaningful gains in English oral language skills for middle school English learners. Frequent teacher demonstrations, vocabulary instruction, and opportunities for students to practice speaking English appear to support student gains. Across the

four years of the study, we asked *DISE* teachers for feedback on the curriculum. Most teachers reported that they enjoyed teaching the program and thought their students responded positively, despite the implementation challenges. They liked the routine and structure of the curriculum for their newcomers, the clear guidance on how to present lessons, and the choral responses. Students liked the predictability of the lessons and the chance to practice hearing and speaking English with other students learning the language.

Future research on *DISE* could focus on other methods for implementation, such as training instructional assistants to teach *DISE* to small homogeneous groups of students in person or remotely to free up teacher time. As of the end of the study, one teacher consistently continued to teach *DISE* remotely, and three teachers periodically taught remotely during the COVID-19 pandemic, suggesting the flexibility of *DISE* for teaching English language skills using other delivery mechanisms.

Declaration of interest statement

The Authors confirm that there are no relevant financial or non-financial competing interests to report.

Acknowledgments

The authors thank Carol Black for making and maintaining the project database, Derek Kosty for his assistance with the observation and coding data sets, and Susan Long for preparation of this manuscript.

Funding

This work was supported by the Institute of Education Sciences under Grant R305A150325.

References

Alemayehu, D. (2011). Current issues with covariate adjustment in the analysis of data from randomized controlled trials. *American Journal of Therapeutics*, *18*(2), 153–157.. 10.1097/MJT.0b013e3181b7d228 doi:10.1097/MJT.0b013e3181b7d228

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology*, *20*, 93–114. doi:10.2307/271083

Allison, P. D. (2009). Missing data. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 72–89). Newbury Park, California: Sage Publications Ltd. doi:10.4135/9780857020994.n4

Allison, P. D. (2012, April 22–25). Handling missing data by maximum likelihood [Paper presentation]. Orlando, FL: SAS Global Forum. Retrieved from http://www.statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf.

Arens, S. A., Stoker, G., Barker, J., Shebby, S., Wang, X., Cicchinelli, L. F., & Williams, J. M. (2012). *Effects of curriculum and teacher professional development on the language proficiency of elementary English language learner students in the central region.* (NCEE 2012-4013). Mid-Continent Research for Education and Learning. Retrieved from https://files.eric.ed.gov/fulltext/ED530839.pdf.

Baker, S. K., Smolkowski, K., Katz, R., Fien, H., Seeley, J. R., Kame'enui, E. J., & Thomas Beck, C. (2008). Reading fluency as a predictor of reading proficiency in low performing high poverty schools. *School Psychology Review*, *37*(1), 18–37. doi:10.1080/02796015.2008.12087905

Baker, S., Lesaux, N., Jayanthi, M., Dimino, J., Proctor, C. P., Morris, J., … Newman-Gonchar, R. (2014). *Teaching academic content and literacy to English learners in elementary and middle school* (NCEE 2014-4012). Washington, DC: National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. Retrieved from the NCEE website: http://ies.ed.gov/ncee/wwc/publications_reviews.aspx.

Ballard & Tighe. (2010a). *IPT II-Oral English Technical Manual, Forms E & F.*

Ballard & Tighe. (2010b). *IPT II-Oral English Test, Form E.*

Ballard & Tighe. (2010c). *IPT II-Oral English Test, Form F.*

Bauman, K. (2017). School enrollment of the Hispanic population: Two decades of growth. Retrieved from https://www.census.gov/newsroom/blogs/random-samplings/2017/08/school_enrollmentof.html.

Borman, G. D., Park, S. J., & Min, S. (2015). The district-wide effectiveness of the Achieve3000 Program: A quasi-experimental study. Retrieved from https://files.eric.ed.gov/fulltext/ED558845.pdf.

Brown, C. H., Wang, W., Kellam, S. G., Muthén, B. O., Petras, H., Toyinbo, P. … Windham, (2008). Methods for testing theory and evaluating impact in randomized field trials: Intent-to-treat analyses for integrating the perspectives of person, place, and time. *Drug and Alcohol Dependence*, *95*, S74–S104. doi:10.1016/j.drugalcdep.2007.11.013

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23–35. doi:10.1007/s00265-010-1029-6

Callahan, R. M. (2005). Tracking and high school English learners: Limiting opportunity to learn. *American Educational Research Journal*, *42*(2), 305–328. doi:10.3102/00028312042002305

Carver, S., & Klahr, D. (2001). *Cognition and instruction: Twenty-five years of progress*. Mahwah: Erlbaum.

Council of Chief State School Officers (CCSSO). (2012). *Framework for English language proficiency development standards corresponding to the common core state standards and the next generation science standards*. Retrieved from https://ccsso.org/sites/default/files/2017-11/ELPD%20Framework%20Booklet-Final%20for%20web.pdf.

Cummings, K. D., Smolkowski, K., & Baker, D. L. (2021). Comparison of literacy screener risk selection between English proficient students and English learners. *Learning Disability Quarterly*, *44*(2), 96–109. doi:10.1177/0731948719864408

Donner, A., & Eliasziw, M. (1987). Sample size requirements for reliability studies. *Statistics in Medicine*, *6*(4), 441–448.

Ellis, R. (2005). Principles of instructed language learning. *System*, *33*(2), 209–224. doi:10.1016/j.system.2004.12.006

Engelmann, S., & Carnine, D. (1991). *Theory of instruction: Principles and applications*. California: ADI Press.

Engelmann, S., Johnston, D., Engelmann, O., & Silbert, J. (2010). *Direct Instruction Spoken English (DISE)*. Colorado: Sopris West.

Estrada, P. (2014). English learner curricular streams in four middle schools: Triage in the trenches. *The Urban Review*, *46*(4), 535–573. doi:10.1007/s11256-014-0276-7

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken: Wiley.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature* (FMHI Publication #231). University of South Florida. Retrieved from https://nirn.fpg.unc.edu/sites/nirn.fpg.unc.edu/files/resources/NIRN-MonographFull-01-2005.pdf.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, *5*(3), 239–256. doi:10.1207/S1532799XSSR0503_3

Geva, E., & Farnia, F. (2012). Developmental changes in the nature of language proficiency and reading fluency paint a more complex view of reading comprehension in ELL and EL1. *Reading and Writing*, *25*(8), 1819–1845. doi:10.1007/s11145-011-9333-8

Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). Institute for the Development of Educational Achievement. Retrieved from http://dibels.uoregon.edu/.

Goswami, U. (2004). Neuroscience, education, and special education. *British Journal of Special Education*, *31*(4), 175–181. doi:10.1111/j.0952-3383.2004.00352.x

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576. doi:10.1146/annurev.psych.58.110405.085530

Graham, J. W. (2012). *Missing data: Analysis and design*. United States: Springer.

Graham, J. W., & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. *The Journal of Applied Psychology*, *78*(1), 119–128. doi:10.1037/0021-9010.78.1.119

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science : The Official Journal of the Society for Prevention Research*, *8*(3), 206–213. doi:10.1007/s11121-007-0070-9

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, *P* values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. doi:10.1007/s10654-016-0149-3

Halle, T., Hair, E., Wandner, L., McNamara, M., & Chien, N. (2012). Predictors and outcomes of early vs. later English language proficiency among English language learners. *Early Childhood Research Quarterly*, *27*(1), 1–20. doi:10.1016/j.ecresq.2011.07.004

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177. doi:10.1111/j.1750-8606.2008.00061.x

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Milton Park: Routledge.

Kim, J. (2011). *Relationships among and between ELL status, demographic characteristics, enrollment history, and school persistence*. CRESST Report 810. Retrieved from https://files.eric.ed.gov/fulltext/ED527529.pdf.

Lawrence, J. F., Capotosto, L., Branum-Martin, L., White, C., & Snow, C. (2012). Language proficiency, home-language status, and English vocabulary development: A longitudinal follow-up of the Word-Generation program. *Bilingualism: Language and Cognition*, *15*(3), 437–451. doi:10.1017/S1366728911000393

Lesaux, N. K., Kieffer, M. J., Kelley, J. G., & Harris, J. R. (2014). Effects of academic vocabulary instruction for linguistically diverse adolescents: Evidence from a randomized field trial. *American Educational Research Journal*, *51*(6), 1159–1194. doi:10.3102/0002831214532165

Low, S., Smolkowski, K., Cook, C., & Desfosses, D. (2019). Two-year impact of a universal social-emotional learning curriculum: Group differences from developmentally sensitive trends over time. *Developmental Psychology*, *55*(2), 415–433. doi:10.1037/dev0000621

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. Oxford, England: Oxford University Press.

National Academies of Sciences, Engineering, and Medicine. (2017). *Promoting the educational success of children and youth learning English: Promising futures*. Washington, D.C: The National Academies Press. doi:10.17226/24677

National Center for Education Statistics (NCES) (2020). *English language learners in public schools*. Retrieved from https://nces.ed.gov/programs/coe/indicator_cgf.asp.

Odom, S. L. (2009). The tie that binds: Evidence-based practice, implementation science, and early intervention. *Topics in Early Childhood Special Education*, *29*(1), 53–61. doi:10.1177/0271121408329171

Raths, J. (1967). The appropriate experimental unit. *Educational Leadership*, *25*, 263–266.

SAS Institute. (2017). *SAS/STAT® 14.3 user's guide*. SAS Institute, Inc. Retrieved from https://support.sas.com/documentation/onlinedoc/stat/142/mixed.pdf.

Saunders, W., Goldenberg, C., & Marcelletti, D. (2013). English language development, guidelines for instruction. *American Educator, Summer*, *13–25*, 38–39. Retrieved from https://www.aft.org/sites/default/files/periodicals/Saunders_Goldenberg_Marcelletti.pdf.

Scammacca, N. K., Roberts, G., Vaughn, S., & Stuebing, K. K. (2015). A meta-analysis of interventions for struggling readers in grades 4-12: 1980-2011. *Journal of Learning Disabilities*, *48*(4), 369–390. doi:10.1177/0022219413504995

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. doi:10.1037/1082-989X.7.2.147

Schrank, F. A., Mather, N., & McGrew, K. S. (2014). *Woodcock-Johnson IV tests of oral language*. California: Riverside.

Scientific Learning Corporation. (2004). *Fast ForWord® language series*. Retrieved from https://www.scilearn.com/program/.

Shaywitz, S., Morris, R., & Shaywitz, B. (2008). The education of dyslexic children from childhood to young adulthood. *Annual Review of Psychology*, *59*, 451–475. doi:10.1146/annurev.psych.59.103006.093633

Shoukri, M. M., Asyali, M. H., & Donner, A. (2004). Sample size requirements for the design of reliability study: Review and new results. *Statistical Methods in Medical Research*, *13*(4), 251–271. doi:10.1191/0962280204sm365ra

Slama, R. B. (2012). A longitudinal analysis of academic English proficiency outcomes for adolescent English language learners in the United States. *Journal of Educational Psychology*, *104*(2), 265–285. doi:10.1037/a0025861

Slocum, T. A., Street, E. M., & Gilberts, G. (1995). A review of research and theory on the relation between oral reading rate and reading comprehension. *Journal of Behavioral Education*, *5*(4), 377–398. doi:10.1007/BF02114539

Smolkowski, K., & Gunn, B. (2012). Reliability and validity of the Classroom Observations of Student-Teacher Interactions (COSTI) for kindergarten reading instruction. *Early Childhood Research Quarterly*, *27*(2), 316–328. doi:10.1016/j.ecresq.2011.09.004

Smolkowski, K., Crawford, L., Seeley, J. R., & Rochelle, J. (2019). Introduction to implementation science for research on learning disabilities. *Learning Disability Quarterly*, *42*(4), 192–203. https://doi.org/10.1177/0731948719851512. doi:10.1177/0731948719851512

Smolkowski, K., Danaher, B. G., Seeley, J. R., Kosty, D. B., & Severson, H. H. (2010). Modeling missing binary outcome data in a successful web-based smokeless tobacco cessation program. *Addiction (Abingdon, England)*, *105*(6), 1005–1015. https://doi.org/10.1111/j.1360-0443.2009.02896.x. doi:10.1111/j.1360-0443.2009.02896.x

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, *60*(2), 263–308. doi:10.1111/j.1467-9922.2010.00562.x

Stanovich, K. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. England: Guilford.

Stein, M., Stuen, C., Carnine, D., & Long, R. M. (2001). Overcoming learning difficulties. *Reading and Writing Quarterly*, *17*(1), 5–24.

Stevens, C., Fanning, J., Coch, D., Sanders, L., & Neville, H. (2008). Neural mechanisms of selective auditory attention are enhanced by computerized training: Electrophysiological evidence from language-impaired and typically developing children. *Brain Research*, *1205*(18), 55–69. doi:10.1016/j.brainres.2007.10.108

Stockard, J., Wood, T. W., Coughlin, C., & Khoury, C. R. (2018). The effectiveness of direct instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research*, *88*(4), 479–507. doi:10.3102/0034654317751919

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP). (2017). *Mathematics and reading assessments*. Retrieved from the Nation's Report Card, April 2018. Retrieved from https://www.nationsreportcard.gov/math_2017/#/nation/achievement?grade=4. and https://www.nationsreportcard.gov/reading_2017/#/nation/achievement?grade=4.

Umansky, I. M. (2016). To Be or Not to Be EL: An Examination of the Impact of Classifying Students as English Learners. *Educational Evaluation and Policy Analysis*, *38*(4), 714–737. doi:10.3102/0162373716664802

Umansky, I. M., Thompson, K. D., & Díaz, G. (2017). Using an Ever–English Learner framework to examine disproportionality in special education. *Exceptional Children*, *84*(1), 76–96. doi:10.1177/0014402917707470

Vaughn, S., Martinez, L. R., Wanzek, J., Roberts, G., Swanson, E., & Fall, A., M. (2017). Improving content knowledge and comprehension for English language learners: Findings from a randomized control trial. *Journal of Educational Psychology*, *109*(1), 22–34. doi:10.1037/edu0000069

Vuchinich, S., Flay, B. R., Aber, L., & Bickman, L. (2012). Person mobility in the design and analysis of cluster-randomized cohort prevention trials. *Prevention Science : The Official Journal of the Society for Prevention Research*, *13*(3), 300–313. https://doi.org/10.1007/s11121-011-0265-y. doi:10.1007/s11121-011-0265-y

Wallace, D., & Green, S. B. (2002). Analysis of repeated measures designs with linear mixed models. In D. S. Moskowitz & S. L. Hershberger (Eds.), *Modeling intraindividual variability with repeated measures data: Methods and applications* (pp. 103–134). New Jersey: Lawrence Erlbaum Associates.

Warren, S. F., Fey, M. E., & Yoder, P. J. (2007). Differential treatment intensity research: A missing link to creating optimally effective communication interventions. *Mental Retardation and Developmental Disabilities Research Reviews*, *13*(1), 70–77. doi:10.1002/mrdd.20139

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. doi:10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$. *The American Statistician*, *73*(Suppl. 1), 1–19. [Editorial]. doi:10.1080/00031305.2019.1583913

What Works Clearinghouse. (2012). *English learners - Find what works based on the evidence.* Washington, DC: U.S. Department of Education, Institute of Education Sciences. Institute of Education Sciences, National Center for Education Evaluation, WWC. Retrieved from http://ies.ed.gov/ncee/wwc/findwhatworks.aspx.

What Works Clearinghouse. (2020). *What Works Clearinghouse standards handbook* (Version 4.1). U.S. Department of Education, Institute of Education Sciences. Institute of Education Sciences, National Center for Education Evaluation, WWC. Retrieved from http://ies.ed.gov/ncee/wwc/.

Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education*, *15*, 345–422. doi:10.2307/1167368

Zvoch, K., & Stevens, J. J. (2006). Longitudinal effects of school context and practice on middle school mathematics achievement. *The Journal of Educational Research*, *99*(6), 347–356. doi:10.3200/JOER.99.6.347-357