AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Educational Psychology

© 2022 American Psychological Association
ISSN: 0022-0663

2022, Vol. 114, No. 7, 1495–1532
https://doi.org/10.1037/edu0000758

Bringing Assessment-to-Instruction (A2i) Technology to Scale: Exploring the Process From Development to Implementation

Carol McDonald Connor[1], Henry May[2], Nicole Sparapani[3], Jin Kyoung Hwang[1], Ashley Adams[1], Taffeta S. Wood[1], Sarah Siegal[4], Cassidy Wolfe[1], and Stephanie Day[1]

[1] School of Education, University of California, Irvine
[2] College of Education and Human Development, University of Delaware
[3] School of Education and the MIND Institute, University of California, Davis
[4] Learning Ovations, Inc., Irvine, California, United States

Bringing effective, research-based literacy interventions into the classroom is challenging, especially given the cultural and linguistic diversity of today's classrooms. We examined the promise of Assessment-to-Instruction (A2i) technology redesigned to be used at scale to support teachers' implementation of the individualized student instruction (ISI) intervention from kindergarten through third grade. In seven randomized controlled trials, A2i and ISI have demonstrated efficacy. However, the research version of A2i was not scalable. To bring A2i to scale in schools serving linguistically diverse students, we carried out the current study across two phases. This study represents both an exploration of what it takes to bring an educational intervention to scale (Phase 1) and a quasi-experiment on the literacy outcomes of learners whose teachers used the technology (Phase 2). We integrated assessments of vocabulary, word decoding, and reading comprehension; revised the A2i algorithms to account for the constellation of skills English learners (ELs) bring to the classroom; updated the user interfaces and added new graphic features; and improved bandwidth and stability of the technology. Findings were mixed, including several nonsignificant results, a marginally significant intent-to-treat effect on word reading in kindergarten and first grade for English monolingual students and ELs, and one significant interaction effect, which suggested ELs and students with less developed reading skills in second and third grade benefited most from the intervention. With some caution, we conclude that A2i demonstrates potential to be used at scale and promise of effectiveness for improving code-focused skills for diverse learners.

Educational Impact and Implications Statement

In this study, we outline the process of bringing Assessment-to-Instruction (A2i) technology to scale within kindergarten through third grade classrooms serving linguistically diverse learners. We carried out this research within two interactive phases. Within Phase 1, we worked closely with our school partners to guide the revision of A2i technology to use at scale. In Phase 2, we conducted a quasi-experiment on the literacy outcomes of learners whose teachers used A2i. Overall, our newly designed A2i technology shows promise to use at scale with kindergarten and first grade monolingual students and English learners. Limitations, implications, and future directions are discussed.

Keywords: individualized or differentiated instruction, instruction, literacy, reading, technology

This article was published Online First September 8, 2022.

Carol McDonald Connor https://orcid.org/0000-0001-5803-456X
Henry May https://orcid.org/0000-0002-5965-3933
Jin Kyoung Hwang https://orcid.org/0000-0003-3074-8047
Ashley Adams https://orcid.org/0000-0002-9960-3199

Ashley Adams is now at School of Speech, Language, and Hearing Sciences, San Diego State University. Taffeta S. Wood is now at the UC Davis MIND Institute, University of California, Davis. Stephanie Day is now at The Readability Consortium, University of Central Florida.

We thank the entire ISI lab team for their contributions to this study with a special thank you to our school partner administrators and teachers, and to the parents and students who participated. Funding to support this project included the U.S. Department of Education, Institute of Education Sciences, Grants R305A160404, R305N160050, and R305A210077, as well as National Institute of Child Health and Human Development P50 HD052120. Carol McDonald Connor has an equity interest in Learning Ovations, a company that may potentially benefit from the research results. The terms of this arrangement have been reviewed and approved by the University of California, Irvine, in accordance with its conflict of interest policies. Sarah Siegal, who was a doctoral student at the time this study was conducted, now works for Learning Ovations, Inc. All results and reporting have been carefully reviewed by authors who do not have a conflict of interest.

Correspondence concerning this article should be addressed to Nicole Sparapani, School of Education, University of California, Davis, One Shields Way, Davis, CA 95817, United States. Email: njsparapani@ucdavis.edu

Moving from research to practice is one of the most difficult challenges confronting practitioners, policymakers, and researchers today. It is critical to make evidence-based technology, programs, professional development, and other materials developed with federal funds accessible to practitioners (Fixsen et al., 2013). Unfortunately, many effective programs developed by researchers sit on shelves or computers. The Department of Education, Institute of Education Sciences (IES) has funded the development and testing of more than 300 programs. Of these, more than 90 programs were efficacious, yet only a small proportion are now used in schools (Albro, 2020). Education is not alone in its challenge to promote the use of evidence-based interventions in communities. Public health, medicine, and other professions share many of the same challenges. These challenges include, but are not limited to, user training and development, cost, and effectiveness at scale. Technology offers additional challenges: user access to technology and Internet bandwidth, feasibility and intuitiveness of design, security, school site positionality toward change, and more.

In many, if not all, applied research studies conducted within the field of education, the goal is to contribute to the body of knowledge within the research community as well as bridge the gap from research to practice in actual classrooms. Closing the gap between research and practice "requires a broader systems perspective that leads to scaled up use of effective practices" (Odom et al., 2020). This bridge becomes tangible with the implementation of technology when considering the number of barriers between controlled research environments to large-scale application (Supplee & Metz, 2015). Hence, this study investigated how we approached and addressed barriers to school-wide implementation of Assessment-to-Instruction (A2i) technology, a web-based literacy tool to support individualized student instruction (Connor, Morrison, Fishman, et al., 2007; Connor et al., 2016), as well as redesigned the technology to be scalable beyond a constrained research setting. We examine A2i as a tool to support literacy development for both monolingual students and English Learners (ELs). This initiative addresses the growing need for programs to effectively meet the needs of today's linguistically diverse student body as well as the increasing call from leading researchers to focus on how to translate decades of reading research, or the "science of reading," to practical implementation by teachers in schools (Solari et al., 2020).

The Present Study—Purpose Statement

The purpose of this effort was to describe the transition from the research version of A2i to a more generalizable platform that contained the needed components vital for improving student literacy outcomes. We had to ensure that the A2i technology had the flexibility and stability for effective implementation in schools nationwide. In the present study, we report aspects of both an exploration of what it takes to bring an educational intervention to scale and a quasi-experiment on the literacy outcomes of linguistically diverse learners whose teachers used A2i. Through this interactive process, we begin to establish evidence of consequential validity of the A2i technology. We present both aspects of scalability within Phase 1 and student level outcomes from the quasi-experiment within Phase 2 together because technology best improves education when it is considered in tandem with student learning rather than on its own (Hantula, 2019). Moreover, implementing at scale includes considering the populations that will be affected by the intervention as it reaches more students within classrooms. For example, ELs are more likely to be reached by an intervention as it spreads to more classrooms. Hence, this article intends to serve as a description of the scalability process while also providing initial evidence of or promise for the effectiveness of A2i at scale.

We begin with presenting the theoretical frameworks that underlie the A2i research technology and briefly outline the features of the tool to provide a foundation for the current project. We then present a model drawn from the implementation science field that we used to guide our process for "scaling up." The project is organized across two phases. Phase 1 is the Exploration Phase (2014–2015). Here, we outline the process and procedures of the exploratory work that provided the foundation for executing Phase 2. We also reflect on lessons learned during the implementation process that allowed us to identify barriers and enact responsive solutions to bringing a revised A2i to scale in kindergarten through third grade classrooms. Phase 2 (2015–2016) is the Quasi-Experimental Phase. Here, we describe our process for developing valid, reliable, and adaptive literacy assessments integrated into the revised A2i technology using a linguistically diverse sample of students. We also present the procedures of and findings from the quasi-experiment. We outline the Method and Results of Phases 1 and 2 separately; however, we interpret our findings from both phases in light of the potential for national scalability.

Theoretical Frameworks Underlying A2i Technology

The theoretical basis for the development of A2i was heavily influenced by the Simple View of Reading (Hoover & Gough, 1990), which outlines the importance of both decoding (code-focused) and language comprehension (meaning-focused) skills for successful reading comprehension. This theoretical model posits that strong code-focused and meaning-focused skills are necessary for reading and comprehending text—without the development of both skills, reading comprehension is jeopardized. There has been extensive empirical evidence supporting the Simple View of Reading not only for monolingual English speakers but also for ELs (e.g., Florit & Cain, 2011; Kim, 2017; Mancilla-Martinez & Lesaux, 2017; Proctor et al., 2006). This justified the recommendations of both code- and meaning-focused instruction provided by A2i for both monolingual English speakers and ELs.

A2i has more recently been informed by the Lattice Model (Connor, 2016; Connor et al., 2016), which places instruction as a central force for change in students' literacy learning. Aligned with Cronbach's (1957) idea of aptitude by treatment interaction effects, the Lattice model emphasizes that the effect of instruction depends on each student's linguistic, text-specific, cognitive, and social-emotional skills (i.e., child characteristic by instruction interaction effects; Connor, Morrison, Underwood, 2007). In other words, the effects of instruction may differ based on students' baseline skills across various developmental domains. Moreover, according to the Lattice Model, there are reciprocal or bidirectional effects such that, as instruction improves literacy skills, it also improves linguistic, cognitive, and social-emotional skills. At the same time, these developmental areas help to improve students' literacy skills (Connor et al., 2016). This idea of students' characteristics (skills) by instruction interaction effects on literacy, as supported by the Lattice Model, are the premise for individualizing student instruction. We next provide a brief overview of A2i.

We refer the reader to Connor (2019) for a full description of the A2i features.

Components of the A2i Technology—Overview of the Research Version

DFI Algorithms and the Classroom View

As supported by the Lattice Model, A2i provides the means for teachers to individualize instruction based on the characteristics that their students bring with them into the classroom, in this case, their literacy skills. At the heart of A2i, and the premise for individualizing student instruction, there are dynamic forecasting intervention (DFI) algorithms. These DFI algorithms are patented (Connor et al., 2013) and developed from empirical studies (e.g., Connor et al., 2004). DFI algorithms compute recommended amounts (in minutes) of four types of literacy instruction that will optimize literacy gains based on individual student's language and literacy skills. The four types of literacy instruction include code-focused instruction with the teacher (e.g., phonological awareness, phonics, spelling, word fluency), meaning-focused instruction with the teacher (e.g., language, vocabulary comprehension, metacognition), code-focused instruction with peers or alone (e.g., phonics worksheets) and meaning-focused instruction with peers or alone (e.g., independent sustained silent reading, buddy reading). With the right information about individual students, teachers can predict students' potential trajectories as they learn to read, taking into account documented sources of influence (e.g., amount of literacy instruction, support from home) and constraints (e.g., previous achievement, home resources). The recommended amounts of instruction are displayed for each student in the *Classroom View* of the A2i technology. As students are assessed throughout the year, the calculated recommendations are automatically updated so that more recent information about students' literacy skills is taken into consideration. The DFI algorithms used in the A2i technology have been tested for efficacy in multiple research studies (Al Otaiba et al., 2011; Connor, Morrison, Underwood 2007, Connor et al., 2009, 2013; Connor, Morrison, Fishman, et al., 2011; Connor, Morrison, Schatschneider, et al., 2011).

A2i Assessments and Graphs

In the research version of A2i, we used standardized reading and vocabulary assessments, administered to students within their schools and entered into the technology by research assistants. Once entered, A2i uses the scores in the DFI algorithms to compute the recommended amounts and types of literacy instruction needed for optimal growth. Each student's assessment results and targeted growth over a one-year period as well as their instructional recommendations are then displayed for teachers within graphs.

Lesson Plans

A2i provides evidence-based resources that teachers can use to individualize instruction based on students' literacy skills. Teachers can access and download (copyright permitting) the activities from their core literacy curriculum and other indexed evidence-based literacy activities (for example, Florida Center for Reading Research [FCRR] center activities; www.fcrr.org). They can also change the activity and locate other relevant activities using advanced search features. Once teachers have given a lesson, they click the activity as accomplished. This records that the activity was completed.

Implementation of A2i Within Kindergarten–Third Grade Classrooms

Although the research version of A2i provided a means for teachers to individualize student instruction, the tool was not feasible nor scalable for classroom use without support from the research team. Previous studies examining the development and effectiveness of A2i have been grounded in design-based implementation research (DBIR)—to develop a tool in collaboration with practitioners that is by design, feasible and implementable (Connor et al., 2015; Fishman et al., 2004). Our aim for this study, however, was that individualizing student instruction, using A2i along with a professional development (PD) protocol, be scalable. In the current paper, we draw from the Exploration, Preparation, Implementation, Sustainment model (EPIS; Aarons, 2011; Moullin et al., 2020) to outline a set of practices and procedures for supporting the implementation of A2i within kindergarten through third grade classrooms with high percentages of ELs. We describe each area in the EPIS model below and contextualize our stages of implementation by drawing from experiences with our school partners across two academic years (2014–2016).

Exploration

Within the EPIS model, the stage of exploration (Odom et al., 2020) takes place at the level of an outer contextual factor (e.g., school districts) and an inner contextual factor (e.g., school administrators; Aarons et al., 2011). In educational settings, these are the district leaders and school principals who make decisions about changes to instruction with which teachers will be tasked. In relation to our project, we met with school principals prior to the start of the study to develop a common research objective. The leaders were tasked with implementing district-mandated Response to Intervention (RTI) within their schools, which included universal literacy screening and multitiered, targeted instruction. Demonstrating how individualizing student instruction with the use of A2i aligned with RTI was the beginning of our mutual partnership, with the shared objective of supporting literacy gains in all learners, including ELs.

Preparation

Schools and teachers possess individual characteristics that vary. During the preparation stage, initial training is provided to site-specific teachers to prepare the climate for implementation, ensuring that schools and teachers have what is needed to create change (Odom et al., 2020). Researchers who work with teachers act as bridging factors or interconnections between research and implementation (Aarons et al., 2011). They must foster trust and "buy-in" of teachers. These teachers, in turn, work with their students to support classroom learning—they act as bridging factors between researchers and students. While this shifting of roles may seem complex, it is in part attributable to the dynamic and reciprocal nature of implementation of change illustrated by the EPIS model (Aarons et al., 2011).

To understand the varying needs and experiences of our school partners, we interviewed school leaders and led workshops with teachers. Our goal was to gather information about the school environment (access to computers and headphones, Internet availability and bandwidth, class size and student characteristics, etc.) as well as individual experiences using technology and running flexible small groups. We used the information learned during this time to prepare the climate for implementation. We then created a roadmap of changes needed for successful scale up. We designed an online professional development (PD) protocol that aligned with the needs of our school partners while also addressing critical components for using A2i to individualize literacy instruction within kindergarten through third grade classrooms.

Implementation

Implementation of an educational intervention positions teachers as learners (Odom et al., 2020). Teachers both provide information and receive feedback on implementation of an intervention, and in turn, use their new learning to change their practice. Fidelity of implementation is critical at this stage as teachers communicate feasibility concerns. In addition, the research team maneuvers or adjusts approaches for different teachers at different stages of "uptake." This might include teachers with different types of experience, degree of openness, and levels of trust that influence intervention implementation. We supported teachers' implementation of A2i through personalized and continuous PD across the school year. We monitored and adjusted our approaches as needed to respond to individual needs, ensure uptake of new practices with fidelity, and facilitate change.

Sustainment

Sustainment can be understood in the context of bringing an educational intervention to scale as the continued implementation of an intervention that has been fully taken up by school sites in classrooms (Odom et al., 2020). Sustainment occurs after researchers have fostered relationships, supported teachers in changing practices, and communicated findings (Aarons et al., 2011). Fostering relationships often begins at the exploration stage and continues throughout the stages. These linkages, as described by the EPIS model, often operate through human and institutional relationships (Aarons et al., 2011). In the case of educational interventions at scale, this would include relationships between teachers and principals, teachers and their students and families, researchers and teachers, districts and researchers, and various combinations of the aforementioned.

At the stage of sustainment, our goal was to give our school partners the tools they needed to continue implementing A2i school-wide without extensive support from the research team, while also maintaining a positive school-researcher partnership. We therefore discussed their progress, shared findings from across the school year, and ensured that everyone (principals and teachers) continued to have access to A2i and the online PD protocol. We also offered continued technical support as needed and an open door for future communication and collaboration.

Phase 1 (2015–2016): Research Objective and Methods

To ensure effective, school-wide implementation of A2i, the primary research objective of Phase 1 was to explore thoroughly the process of scaling up. That is, we examined the transition between implementing the research version of A2i to a more generalizable tool. In Phase 1, we recruited 24 kindergarten through third grade teachers and four principals (one per school site) from two large schools in Phoenix, Arizona (AZ) with substantial EL student populations and two schools in Pittsburg, Pennsylvania (PA).

Procedures. At the start of the academic year, we carried out in-person structured interviews with the school principals from each site to gather information on the individual needs of their schools and establish a reciprocal school-researcher partnership. We inquired about district-level and school-level concerns and noted areas for potential collaboration. Although the schools were tasked with different district-level charges, they shared the common goal of improving literacy outcomes in their early elementary students. We developed a year-long plan for partnership centered on implementing A2i in kindergarten through third grade classrooms to support individualized literacy instruction, while studying the process and gathering feedback from teachers. The schools shared their beginning and end of year progress monitoring data (i.e., DIBELS), and the research team uploaded the scores to A2i per classroom.

Initial Trainings. The school year started with a "kick-off" in-person training for teachers at each school site. The training consisted of two half-day workshops in which we gathered information about the school implementation climate and the needs and experiences of individual teachers and grade-level teams. We also provided information regarding A2i as an evidence-based literacy tool, discussed the features of the research version, and assisted teachers in using A2i in their classrooms to individualize student instruction.

Monthly Communities of Practice Meetings. In addition, two classroom educators from our research team facilitated monthly grade-level communities of practice meetings (e.g., Bos et al., 1999) at the AZ school sites only, because these schools were local to the research team. We developed a working handbook, which included guiding questions and monthly topics (setting up your classroom, using A2i recommendations to drive instruction) to structure the meetings and facilitate discussion. The monthly meetings followed a similar sequence across the schools and grade-levels, including a "check-in" period to inquire about strengths and concerns with individualizing instruction using A2i, delivery of content, and discussion with reflection.

Classroom Observations. In addition to these monthly communities of practice, the classroom educators from our research team observed each of the AZ teachers in their classrooms three times during the year (fall, winter, spring). Specifically, we were interested in understanding whether and how teachers effectively used A2i to plan and deliver literacy instruction within individualized, small groups and differentiated learning centers for their diverse student body. We assisted teachers as needed in understanding the A2i recommendations, creating individualized small groups and learning centers based on the A2i recommendations, and preparing the A2i recommended curricula materials and evidence-based activities.

Focus Groups. Finally, we carried out focus groups with teachers from each site to gather information on their experiences using A2i in their classrooms. For the AZ schools, the teachers, research team, and program developers participated in focus groups (one focus group per site). In the PA schools, the research team met with teachers, gathering notes to share with the program developers at a later time. The focus group questions centered on teachers' experiences with specific features of A2i. We inquired, for example, about the A2i features teachers found most helpful and how easily they were able to navigate the tool as well as readability of tables and figures and usefulness of the A2i recommended materials and activities. This information was critical, because it helped to inform the updates we made to the A2i technology prior to Phase 2.

Data Sources. We collected detailed notes from the initial planning meeting with the school principals, the "kick-off" training, and the monthly communities of practice meetings with our AZ schools. We compared notes from the monthly communities of practice meetings across groups to outline similarities and differences between the different grade levels and schools. In addition, we gathered field notes during the classroom observations and monitored teachers' usage of A2i to support their students' learning as a means for gauging fidelity. Finally, we iteratively reviewed the records taken from the focus groups, in which we elicited teachers' feedback about their experiences using A2i. Taken together, we identified four themes that we addressed prior to the quasi-experiment carried out during the 2015–2016 school year. We next outline barriers and solutions derived from the four themes. See Table 1 for a summary of this process.

Barriers and Solutions to Implementation—Redesigning A2i Technology

Barrier and Solution 1, Effort From Research Team and Integrated Assessments

Perhaps the most daunting barrier identified was the high level of effort required from the research team to administer, score, and enter the assessments that allow the A2i algorithms to make instructional recommendations for individual students. As a result, we determined that A2i would need integrated assessments that students could take with relatively little teacher intervention. We realized that the assessments would need to be short enough for students to take multiple times in a school year, and they would need to provide reliable, valid estimates of students' language and literacy skills. The assessments would also need to be scored automatically, without researcher support. With this in mind, we developed three adaptive assessments validated for students in kindergarten through third grade that could be integrated into A2i: an online vocabulary assessment (Word Match Game [WMG]) and two reading assessments (Letters to Meaning [L2M] and Reading to Comprehension [R2C]). Details on item development and psychometric properties are reported in Table 1 and in the Method section.

Barrier and Solution 2, User Interface and Improved Lesson Plans

The second barrier was related to the user's experience of the user interface (i.e., how easy A2i was to navigate and use). Teachers and administrators reported wanting additional information about the lesson plans, specifically how they related to the Common Core State Standards (CCSS; Common Core State Standards Initiative, 2010) and better tools to visualize teacher usage of A2i and student progress across the school year. To be responsive to these requests, we improved and expanded the lesson planning feature, which was used to facilitate automatic lesson planning for the implementation of individualized instruction in the classroom. Specifically, we included search and navigation menus, a wider curriculum selection, indexed curriculum activities linked to the CCSS, and recommended open-source materials linked directly to the lesson plans. We also included enhanced reports for student progress and teacher usage, improved reporting features as well as added more web-based PD resources. See Table 1 for details and Appendix A for screenshots.

Barrier and Solution 3, Recommendations for ELs and Updating the A2i Algorithm

A third theme that emerged from the data was teachers' desire to understand how to interpret the A2i recommendations for ELs. The initial studies that demonstrated the efficacy of A2i were conducted in areas that had a diverse cultural and racial makeup, but they were not diverse linguistically. Considering the growing number of ELs attending elementary school in the United States, and the fact that the teachers involved in Phase 1 of the study were in AZ and PA, it is not surprising that this issue arose. Having an intervention that scales up means having an intervention that works for all students, including students from culturally and linguistically diverse backgrounds.

Although scholars of effective instruction for ELs call for more research on modifications to classroom instruction for ELs, they have identified several strategies that are advantageous to literacy development including, individualizing (or differentiating) instruction (Gunn et al., 2000; Kamps et al., 2007), providing ongoing teacher support and student monitoring (Haager & Windmueller, 2001), identifying similarities and differences between students' first and second languages (Giambo & McKinney, 2004; Kramer et al., 1983), and capitalizing on first language strengths (August et al., 2014; August & Shanahan, 2010). A number of classroom-level intervention studies that have focused on ELs have also shown positive effects in enhancing students' language and literacy skills (e.g., Calderon et al., 1998; Cheung & Slavin, 2012; Collins, 2014; Vaughn et al., 2005). Drawing from this evidence and from the Simple View of Reading framework, we concluded that individualizing instruction using both code- and meaning-focused instructional recommendations from A2i would be appropriate for ELs, but we considered the need to revise the A2i algorithms to accommodate ELs' unique constellations of skills.

Given that the integrated A2i assessments were developed to measure literacy skills in English, we reevaluated the appropriateness of the algorithms to make instructional recommendations for ELs (who were receiving English-only instruction) based on their current literacy skills in English. The information that feeds the algorithm for recommendations related to time spent in meaning-focused instruction is pulled from student performance on the vocabulary assessment (for kindergarten and first grade) and from the reading comprehension assessment (for second and third grade). ELs with limited oral language proficiency in English would be expected to score lower than children with higher levels of English oral language proficiency on these assessments, which would lead the algorithms to recommend more time in teacher-managed, meaning-focused instruction. Increased time in small-group instruction that supports oral language development aligns with recommendations within the existing literature related to how best to support ELs in the classroom (e.g., August et al., 2016, 2018; Baker et al., 2014; Crevecoeur et al., 2013; Gersten & Baker, 2000; Gunn et al., 2000; Shanahan & Beck, 2006). We recognize, however, that more precise recommendations could likely be made by incorporating both English and native language skill—this is a direction for future work.

Table 1
Summary of the Procedures and Key Points Outlined in Phase 1

Identified barrier and data sources that informed decisions	Solution	Evidence to support scalability	Key points and recommendations
Barrier 1, Effort from Research Team: The high level of effort required from the research team to administer, score, and enter the assessments that allow for the A2i algorithms to make their instructional recommendations for each individual student. We documented the amount of time that the research team spent on gathering assessment information and uploading scores into A2i.	Solution 1, Integrated Assessments: We developed and tested three literacy assessments that were integrated into A2i. The assessments are adaptive, so students begin each assessment at their grade level, but the difficulty level of the items either increases or decreases based on the students' performance. For example, if students miss an item, the next item is easier; if they get the item correct, the next item is more difficult. This allows for relatively quick administration of each assessment (approximately 7–10 minutes per assessment). The assessments also provide a reliable and valid measure of students' literacy skills, which are used in the A2i algorithms to make instructional recommendations for every student. The test results and instructional recommendations are updated in real time with each completed assessment	With the redesign, the research team did not need to collect assessment data or upload scores into A2i. Teachers were able to administer the assessments independently with some assistance from the research team as needed. As a result of the adaptive nature of the assessments, students were able to take the online assessments throughout the year without seeing the same items multiple times. Teachers were able to monitor and track their literacy progress over time and make changes to their practices based on the assessment information. These assessments were further improved to be functional on iPads and Tablets, which improved the flexibility of use for schools and the reliability of the scores for younger students.	Exploration. Centering a school–research partnership on a common goal is critical for successful implementation of school-based interventions. Universal screening, effective progress monitoring, and targeted tiered instruction that leads to literacy achievement in all learners are the school- and district-level objectives that provided the entry point for our study. By establishing a mutual partnership, in which school leaders and teachers were key players in our study, we were able to successfully redesign and implement A2i within classrooms—ensuring that A2i provided teachers the means to monitor their students' literacy progress and make instructional decisions with ease.
Barrier 2, User Interface: Teachers wanted additional information about the lesson planning feature, lesson plans to link with Common Core State Standards (CCSS; Common Core State Standards Initiative, 2010), and better data visualization of student progress. School and district leaders wanted reports on how A2i was being used in individual classrooms. We iteratively reviewed records taken from the focus groups and revised the user interface based on teachers' feedback and suggestions.	Solution 2, Improved Lesson Plans: We added search and navigation menus, a wider curriculum selection, and indexed curricula materials that were linked with CCSS. A set of administrative menus were also added, allowing new curricula and resources to be added directly to the A2i lesson database. New curricula materials and resources continue to be indexed and stored in the A2i *Lesson Plan*. Student progress reports were enhanced, and teacher usage reports and tracking features (tracking user-clicks per page visits) were included to facilitate district and school educational leaders' provision of focused support to teachers for individualizing student instruction.	These updates improved the flexibility of the program and expanded the number of activities teachers could address to meet their students' diverse learning needs overall. Teachers were able to independently navigate the A2i features and pages and implement the recommended activities that aligned with their curriculum, which were linked to the CCSS. Reporting features allowed teachers and school administers to access and export student test scores, making these data easily available at both the school and district level. School leaders were also able to review and download teacher user logs, which provided the amount of time that teachers used the varying A2i features.	Sustainment. Fostering positive relationships among school leaders, teachers, students, and the research team is foundational for sustainability. A2i needed to be an accessible, flexible, reliable, and stable tool that provided teachers with the information they needed, in a format they could read and interpret, to individualize student literacy instruction. With easy access to teacher usage and student progress information, A2i provided a platform for communication between school leaders and teachers. This was an important component for sustainability, as school leaders are key players in supporting teachers in changing their practices.
Barrier 3, Recommendations for English Learners (ELs): Teachers wanted to know how to interpret the A2i recommendations for ELs. With the growing number of ELs attending elementary school in the United States, scaling up meant that A2i needed to work for all students, including students from culturally and linguistically diverse backgrounds. During the in-person, structured interviews with school principals, the initial "kick-off" training" and the classroom observations, we learned of the district-wide goal	Solution 3, Recommendations for ELs: We revised the algorithms to make instructional recommendations for ELs based on their current literacy skills in English. Because our sample of ELs demonstrated limited vocabulary skills, we included vocabulary in the A2i algorithm. By doing so, A2i provided recommendations for both teacher-managed meaning-focused time and additional teacher-managed code-focused time based on students' vocabulary skills.	For children in the United States who speak a language other than English at home, research has documented that well-designed education programs with appropriate assessments can successfully support their achievement in both English and their home language (Bialystok, 2001; Collins, 2014; Francis et al., 2006). This appears to be especially the case for ELs who speak Spanish at home (e.g., Baker et al., 2016; Collins, 2014). Using the revised the algorithms, teachers are able to deliver individualized literacy instruction to	Implementation. Positioning teachers as learners and ensuring interventions are carried out with high fidelity are critical steps in the implementation process. We revised the A2i algorithms to ensure that teachers were able to use A2i to make data-driven instructional decisions for all their students, including ELs. We provided teachers with personalized PD throughout the year to gather feedback on the implementation process and provide them with information to move forward. These meetings also helped us gauge whether teachers were using A2i

(*table continues*)

Table 1 (*continued*)

Identified barrier and data sources that informed decisions	Solution	Evidence to support scalability	Key points and recommendations
of improved literacy outcomes in all learners, including ELs.		all of their students, including ELs with varying levels of English proficiency. This is especially critical in districts serving large numbers of culturally and linguistically diverse students.	as a tool to individualize student instruction, which in part, provided us with a measure of fidelity.
Barrier 4, Bandwidth: Slower-than-normal response times from the website during times when website traffic was high. We became aware of this issue during the communities of practice meetings and the classroom observations when assisting teachers in using A2i.	Solution 4: The infrastructure of the servers, codebase, and internal data tables were enhanced. The capacity to handle large numbers of simultaneous users greatly improved the flexibility and power to the technology. In addition, purely logistical implications of scale included increased data security and capacity needs within the system. Security updates were made to the website as well as the password system to allow for higher levels of data security.	Teachers were able to use A2i without response times slowing during high traffic times. Multiple teachers within schools were able to access the varying A2i features simultaneously, and students were able to access the web-based A2i assessments directly without having to navigate A2i. The capacity to handle large numbers of simultaneous users greatly improved the flexibility and power of the technology.	Preparation. Preparing the climate for successful school-wide implementation is foundational. We needed to ensure that the environment was adequately equipped to support change. To benefit from A2i, we learned that (a) we needed teachers' "buy-in" about the tool, (b) teachers needed to have access to multiple computers (at any given time), and (c) school sites needed to have fast and reliable internet connectivity. Issues at this stage of the implementation process could have jeopardized successful implementation efforts overall.

Note. A2i = Assessment-to-Instruction.

When considering the A2i algorithm's recommendations for teacher-managed, code-focused instruction, we explored whether to base this recommendation solely on word reading skills (as had been the case with previous A2i studies among English-only students) or to include vocabulary scores so that students with lower levels of vocabulary would receive recommendations for larger amounts of teacher-managed, code-focused instruction. Our rationale for ultimately altering this algorithm to include both word reading and vocabulary skills was that students with less developed English vocabularies would benefit from spending relatively more instructional time with the teacher where they would be most likely to receive explicit, code-focused instruction tailored to their individual needs. Again, we based this conclusion on theory as well as the literature related to best instructional practices for ELs (e.g., Baker et al., 2014; Cunningham & Stanovich, 1997; Ouellette, 2006; Perfetti & Hart, 2002; Scarborough, 2001; Thomas & Sénéchal, 2004). See Table 1 for additional information and further rationale.

Barrier and Solution 4, Bandwidth

The final barrier was identified as a result of teacher reports of occasional slower-than-normal response times from the website, which the research team identified as being related to times when website traffic was high. To address the increase in traffic inherent in scale up, the infrastructure of the servers, codebase, and internal data tables were enhanced to account for additional users without reducing performance. To reduce traffic on the main website, a protocol was also developed to enable students to access the online assessments directly, without having to navigate A2i. See Table 1 for further detail.

Phase 2 (2015–2016): The Quasi-Experimental Phase—Research Objectives

Phase 2 aimed to test whether our revised, scalable version of A2i demonstrated promise of effectiveness when implemented by elementary school teachers serving both English monolingual students and ELs. There were three research questions in this quasi-experiment.

1. What is the validity of the newly developed, integrated A2i assessments that are embedded within the A2i technology?

2. What effect does teachers' use of the revised A2i technology, with on-going professional development (PD), have on students' literacy outcomes (intent-to-treat)?

   - Does the effect of A2i depend on students' initial language and literacy skills?
   - Does the effect of A2i depend on whether students are monolingual or EL?

3. Controlling for preintervention reading scores, are postintervention reading scores higher for those students whose teachers spent more time using the A2i technology?[1]

   - To what extent does teachers' use of the revised A2i technology, calculated from user logs (treatment teachers only), predict students' reading outcomes?

[1] It is important to note that analyses for research question 3 are exploratory and do not provide support for causal inference about program impacts owing to the likelihood of unmeasured confounds that correlate with both teachers' use of A2i and student outcomes. As such, any significant findings here would suggest that teacher use of A2i is correlated with students' reading gains, but we cannot rule out the possibility that the difference in reading gains is due to an unmeasured confound instead of the impact of A2i.

- Does this vary by students' monolingual or EL status?
- Is teacher use of A2i related to PD uptake?

Method

Transparency and Openness Statement

This research was conducted following a grant proposal funded by the Institute of Education Sciences (IES; Grant # R305A160404), which prespecified the research questions, theoretical framework, implementation strategy, data collection, and analysis plan. As an IES Development Grant, there was no requirement for public release of data, and the IRB protocol and consent forms for this study do not allow for sharing data with third parties. Data analyses were conducted using HLM7 and SAS 9.4; data analysis code is available from the authors on request. Selected materials from the study (e.g., the implementation fidelity rubric) are also available from the authors on request. A2i is now a commercial product, and the authors include a conflict of interest statement printed elsewhere in this article.

Procedure

During the 2015–2016 academic year, we conducted a quasi-experiment to assess the promise of the effectiveness of using A2i to support teachers as they individualized their students' literacy instruction. Two large schools in AZ were randomly assigned to either use A2i at the beginning of the school year (immediate treatment) or to wait until April of the school year (delayed treatment). The school year for both schools ended in June. Both schools used the same curriculum: Wonders, published by McGraw Hill (www.mheducation.com/prek-12/program/microsites/MKTSP-BGA10M0/wonders.html). The Wonders curriculum was indexed (embedded within A2i) so that teachers could access recommended lessons from the A2i Lesson Plan based on their students' grade level and reading ability.

Participants

Thirty-three kindergarten through third grade teachers and their students (N = 763) participated in the quasi-experiment. There were four or five classrooms per grade level at each school. Sixty-eight percent (68%) of the participants qualified for the U.S. National School Lunch Program (NSLP), which is frequently used as a proxy for socioeconomic status. Eighty percent (80%) of the students were Hispanic/Latinx, with 25% designated as ELs. In this district, students identified as nonproficient English Learners (ELs) were assigned to an English immersion classroom (EL classroom), with one EL classroom per grade level per school. EL classrooms had a dedicated four-hour English language block to support English language development. This four-hour block was at the academic expense of other content areas, with mathematics as the exception.

Professional Development

All participating teachers across both treatment conditions received professional development (PD) delivered by educators (certified teachers or classroom specialists) on the research team. However, the PD protocol varied by treatment condition. The teachers in both conditions participated in two half-day workshops prior to the beginning of the school year, but only the immediate treatment condition was given access to A2i at this time. With access to A2i, they were able to access the online PD materials and use all of the A2i features (Lesson Plan, Classroom View, etc.). In addition, the teachers in the immediate treatment condition received personalized coaching in the classroom three times per year and monthly grade-level communities of practice meetings. In the delayed treatment condition, teachers were given access to A2i starting in April.

Measures

Students were administered a battery of well-established, valid, and reliable standardized literacy measures as well as the A2i online literacy assessments. For both conditions, all assessments, excluding the A2i online assessments, were administered in the fall (between August and September depending on classroom schedules) and again in the spring (April). Students in the immediate treatment condition completed the A2i online assessments in the fall and spring; Students in the delayed treatment condition completed the A2i assessments only in spring, just before their teachers began using A2i since accessing the assessments required access to A2i. The spring assessment scores represent the outcome measures for the quasi-experiment. In addition, as a measure of implementation fidelity, we monitored teachers' A2i usage through user-logs and gauged teachers' PD uptake using a researcher-developed rubric.

Standardized Literacy Measures

Woodcock-Johnson III Test of Achievements. The Woodcock-Johnson III Test of Achievements (WJ-III; Woodcock et al., 2001) is a standardized assessment, normed on a nationally representative sample that measures a wide range of students' cognitive and academic abilities. The Letter-Word Identification subtest (LW) was used to assess kindergarten and first graders' ability to name and decode words out of context. Research personnel administered the LW subtest individually to students in a quiet area outside the classroom. Reliability on the subtest in the students' age range varied from .93 to .98. The intraclass correlation (ICC) for the spring assessment was .40, suggesting that 40% of the variability in students' scores fell between classrooms. W scores were used in the analyses.

Gates-MacGinitie Reading Test. The Gates-MacGinitie Reading Test (GM; MacGinitie & MacGinitie, 2006) is a standardized reading assessment that has two subtests: Vocabulary and Reading Comprehension. Research personnel administered the assessment to second and third graders as a whole group within their classrooms. Reliability coefficients ranged from .64 to .75, and the ICC for the spring assessment was .26. Extended scale scores from both subtests were used in the analyses.

A2i Online Assessments

One key aim of this study was to develop integrated online and computer adaptive tests that were valid (i.e., demonstrating both construct and predictive validity) to use in the A2i technology. The use of computer adaptive assessments within A2i allowed for shorter test administration times, with initial item selection determined by a student's grade level and subsequent item selection determined by a student's performance on previously administered items. This maximizes both the efficiency and reliability of the A2i assessments by presenting students with only a subset of items specifically aligned with their current ability level. Sample practice items are presented in Appendix A. Three A2i online assessments, outlined below, were used in the current study: Word Match Game (WMG), Letters2Meaning (L2M), and Reading2Comprehension (R2C). Teachers administered the assessments to students in their classrooms with assistance from the research team as needed.

Word Match Game. This assessment was designed to measure students' vocabulary knowledge using a semantic matching task. Students are presented with three words by audio and text (e.g., cat, kitten, tree). The words are highlighted as they are presented, and students are asked to click two words that go together (e.g., cat and kitten). The assessment is adaptive, requiring students to match more advanced vocabulary words (e.g., copal and resin) if they continue to match correctly or conversely, presenting more simple vocabulary if they are not semantically matching words.

Letters2Meaning. This assessment was designed to assess students' decoding, word reading, spelling, and sentence writing skills (generative comprehension skills and grammatical knowledge). L2M has five consecutive components ranging from simple alphabetic principle tasks to sentence-level semantics: Letter Identification, Letter-Sound Identification, Word Identification, Letters2Words, and Words2Sentences. The easiest task is Letter Identification in which they click on the letter that they hear from a pool of letters. In the Letter-Sound Identification task, students hear a letter sound and are asked to click on the letter that corresponds to the sound from a pool of letters. In the Word Identification task, students are asked to click on the word they hear from a pool of words. In the Letter2Words task, students hear a word and are asked to select letters from a pool of letters to spell out the word. Finally, the Words2Sentences task asks students to create meaningful sentences from a pool of words. Text structure (e.g., punctuation) are included as clues for creating the sentence. This assessment advances through all five components as students answer correctly. The ICC for the April assessment of L2M was .57.

Reading2Comprehension. This assessment was designed for students who read at a second-grade level or higher. R2C measures students' higher-order reading comprehension skills (inferencing and comprehension monitoring) across social studies, science, and narrative text. Students read a passage that is missing a word early in the paragraph and select one of four words to fill in the blank. All four choices make sense when they are first read in the sentence, so students cannot identify the correct word until they read and comprehend the entire paragraph.

Teacher Involvement Measures

A2i-Generated Teacher User-Logs. A2i automatically generates user-logs outlining the amount of time teachers spend using the varying A2i features (e.g., Classroom View, Lesson Plan, etc.). The user-logs can be viewed as charts within A2i for teachers to see, and they can be exported as Excel spreadsheets. For this study, we focused on the total amount of time teachers in the immediate treatment condition used A2i (min), not including the time students spent on assessments.

Teacher PD Uptake Rubric. This researcher-developed rubric included eight items (outlined in Appendix B). One of the educators from the research team rated teachers' PD uptake, ranging from 1 (*poor*) to 5 (*strong*) based on teachers' attendance in the monthly communities of practice meetings, participation in the PD opportunities, and willingness to learn and use A2i within their classrooms.

Psychometric Analyses Plan

Item Response Patterns and Missing Data

Given that items on some A2i assessments were administered via a computer adaptive testing (CAT) platform, which selects items to be administered based on correct/incorrect responses, the specific set of items administered to one student was generally different from the set of items administered to other students. Methods for handling missing data across the full set of items included the EM algorithm to estimate the item covariance matrix and full-information maximum likelihood (FIML) estimation of scaling model parameters and person scores.

Dimensionality

The number of constructs captured by each instrument was examined via exploratory factor analysis and scree plots of factor eigenvalues. The EM algorithm was used to estimate the item covariance matrix given the missing data associated with computer adaptive administration. Data for each instrument were analyzed separately to assess the strength of a single latent construct (i.e., an overall scale) and search for evidence of potential subscales for each instrument. A large eigenvalue for the first factor relative to the second eigenvalue (e.g., $\xi_1 > 3 \times \xi_2$, or $\xi_1 > 2 \times \xi_2$ and $\xi_2 < 1.5$) was considered evidence of unidimensionality.

Scaling Model and Estimation

A Rasch scaling model was used to estimate item difficulty parameters and person scores for each instrument. The Rasch model is an item response theory (IRT) model that expresses the probability of a correct item response as a function of an item's difficulty (b_i) and the respondent's ability (θ). A unidimensional model was used for each assessment. The functional form of the model is:

$$P(Y_i = 1) = \frac{e^{\theta - b_i}}{1 + e^{\theta - b_i}}. \tag{1}$$

The scaling model used estimates item and person parameters using all of the available data and accommodates the differences in item sets administered to different students. The Rasch scaling models were estimated using PROC IRT in SAS STAT 14.1 under SAS 9.4. All items from an instrument were included in the estimation for that instrument, with nonadministered items having missing values for responses. FIML was used to estimate the item parameters based on the complete set of observed item responses, with nonadministered items excluded from likelihood calculations. An item difficulty parameter and its standard error was estimated for each item for which there were at least 30 responses, including at least one incorrect and one correct response. The percent of correct responses was also calculated for each item. Goodness of fit for each item was assessed using Pearson's χ^2 statistic based on the subset of students responding to the item. *p* values were calculated for each item, with values less than .05 suggesting poor fit under the Rasch model.

An overall test information function (TIF) was calculated for each instrument based on the estimated item parameters and associated item characteristic curves. A plot of the TIF curve for each instrument was used to assess the precision of score estimation throughout the range of possible test scores. Information values greater than 2 (i.e., corresponding to a reliability greater than .70) were considered adequate for precise score estimation at that point on the ability scale. Values less than 2 were considered as suggesting the need for additional items with difficulty near that point on the ability scale.

Respondent Scores

An overall score (θ) was estimated for each respondent as the maximum a posteriori (MAP) score, which is equal to a weighted combination of the maximum likelihood (ML) score and a standard normal (*M* = 0, standard deviation = 1) Bayesian prior distribution. MAP scores are highly correlated with ML scores, but they are less prone to problems of estimation and outlier scores for those students who answer most or all items presented to them correctly or incorrectly (e.g., ceiling or floor effects).

Grade equivalent (GE) scores were calculated by linking scores on each A2i assessment to scores on the LW and GM reading assessments administered at approximately the same time (generally within 2–4 weeks) of the A2i assessment. Linking to the standardized assessments allows estimation of GE scores relative to a nationally representative sample of elementary students. Nonlinear regression models using a logit transformation were estimated to determine a conversion equation between the standardized test scale scores and grade equivalents.

Statistical Model of Impacts

A hierarchical linear model (HLM) was used to analyze differences in students' scores from April, 2016 on the A2i Letters2-Meaning test (grades K-3), the LW test (grades K-1), and the GM Reading test (grades 2–3). The mathematical form of the model is:

Level-1 Equation:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(Pretest_{ij}) + r_{ij} \tag{2}$$

Level-2 Equations:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(A2iIMM_j) + \gamma_{02}(Grade_j) + u_{0j} \tag{3}$$

$$\beta_{1j} = \gamma_{10}, \tag{4}$$

with Y_{ij} representing the April test score for student *i* from classroom *j*, $Pretest_{ij}$ representing the fall pretest score for student *i* from classroom *j* (included only for a subset of LW and GM models given that A2i fall scores do not exist for the delayed intervention group), $A2iIMM_{ij}$ indicating whether the class was in the treatment condition ($A2iIMM_j$ = 1) or the delayed treatment (control) condition ($A2iIMM_j$ = 0), and $Grade_j$ indicating the grade level of classroom *j* (K = 0, First = 1, etc.). The coefficients represent the fitted mean score in kindergarten (γ_{00}), the impact of A2i (γ_{01}), an overall grade effect (γ_{02}). Additional parameters are included in some models to estimate the moderating effect (i.e., interaction) of Grade, EL status, or baseline literacy scores on the A2i impact (γ_{11}). Given that this study is focused on feasibility of implementation and potential impacts of A2i as a new intervention, and that it involves a relatively small sample, we do not implement a strict .05 cutoff for significance, nor do we implement a correction for multiple tests. Although this does increase the possibility of a type I error, the exploratory nature of this study calls for more focused control of type II errors.

Results

Research Question 1: The Validity of the A2i Assessments

Results for each of the three assessments based on the series of psychometric analyses described above are summarized in Table 2. Additional details and figures are provided in Appendix C. Results for two of the three A2i assessments, WMG and R2C, suggest unidimentionality, whereas results for L2M suggests a strong general factor, and three to six subscales. Item fit statistics were good for all but a few items, and item difficulty statistics and test information plots suggest adequate reliability of measurement (i.e., I $>$ 2.0, r $>$ .70) throughout a wide range of abilities for both the L2M and WMG computer adaptive tests, whereas the R2C item difficulty statistics and test information plots suggest that the R2C assessment (which does not use CAT) is not appropriate for students with less-developed reading skills.

We reviewed how well each A2i assessment correlated with itself and with standardized measures, including the LW subtest on the WJ-III and the GM. Results are provided in Table 3. The L2M correlated highly with both the LW subtest (given only to kindergarten and first grade students) and the GM (given only to second and third grade students) with correlations (*r*) ranging from .65 to .76. The WMG was moderately correlated with L2M (*r* = .56), although it had smaller correlations with LW and GM (*r* ranging from .27 to .37) and no significant correlation with R2C. R2C was moderately correlated with L2M and to the GM (*r* = .30).

Research Question 2: Effects of A2i on Students' Literacy Outcomes (Intent-to-Treat Results)

Analyses revealed no significant differences between conditions on the standardized measures at baseline, which is required for a strong quasi-experiment (Shadish et al., 2002). Table 4 shows means and standard deviations for treatment and control groups on WJ-III LW subtest and GM assessments at baseline. On average, students were reading at grade expectations in kindergarten and first grade, based on examination of LW standard scores (*M* = 98). However, in second and third grades, based on GM percentile rank, on average students were reading below grade expectations at the 34th percentile. In general, second and third grade students in EL classrooms tended to have lower GM scores (23rd percentile) compared with their peers in general education classrooms. There was no significant difference in LW scores for ELs in

Table 2
Summary of Psychometric Analysis Results

Assessment	Results of IRT analyses
Word Match Game	209 items with more than 30 responses for each Proportion correct across items = .61 Item difficulty ranged from −3.3 to +3.5 (mean difficulty = −.38) Appears to be unidimensional Overall test information was excellent with a bell-shaped function and total information greater than 2.0 throughout the range of Rasch theta scores from −5.0 to +5.0, suggesting that computer adaptive administration of WMG will produce reliable individual scores throughout the full range of student abilities
Letters2Meaning	2,807 test administrations with a majority of students responding to more than 10 items. 505 individual items with more than 30 student responses were used in the IRT analyses. L2M may not be purely unidimensional with a large first factor. Thus, there exists the potential for subscales. Overall test information for the complete pool of 505 Rasch-scaled L2M items was excellent, with a bell-shaped information function and Total Information greater than 2.0 throughout the range of Rasch theta scores from −5.0 to +5.0, suggesting that computer adaptive administration of L2M will produce reliable individual scores throughout the full range of student abilities.
Reading2Comprehension	All 10 items in the R2C item pool had more than 30 responses and were included in the Rasch analyses. The average proportion correct across the items was .37 and the median proportion correct was .32 across all items. Item difficulty parameter estimates for the 10 items ranged from −1.5 to +2.3 with a mean difficulty of +1.28, a median difficulty of +1.53, and a standard deviation of 1.03 points on the Rasch Theta scale. Standard errors for the difficulty estimates ranged from 0.07 to .16 with a mean standard error of 0.11 and a median standard error of 0.10 points on the Rasch Theta scale. Overall test information for the complete pool of 10 Rasch-scaled R2C items was modest, with a bell-shaped information function and Total Information greater than 2.0 for Rasch theta scores in the range +1.0 to +3.0, suggesting that computer adaptive administration of R2C will produce reliable individual scores only in the upper range of student abilities and that reliability of R2C scores at the lower end would be improved if additional items were added to the R2C item pool.

Note. IRT = item response theory; L2M = Letters2Meaning; R2C = Reading2Comprehension; WMG = Word Match Game.

kindergarten or first grade EL classrooms compared with their peers in general education classrooms (Table 4).

Intent-to-Treat Results

Using hierarchical linear modeling (HLM), with students nested in classrooms and a fixed effect denoting school/treatment assignment at the classroom level, we examined intent-to-treat (ITT) effects using the A2i integrated assessments as well as the WJ-III LW subtest for kindergarteners and first graders and the GM for second and third graders. When using the A2i assessments as outcome measures, we found a marginally significant ITT effect on L2M in kindergarten (p = .09) with a small effect size (d = .15). This effect decreased as grade increased (see Table 5). Scores on L2M were higher for students in later grades, but there was no grade by treatment interaction effect. When we added EL classroom to the classroom level of the model, we found a treatment by EL classroom interaction effect (p = .048)—there was generally little effect of treatment for students in general education classrooms, but there was a greater treatment effect for students in English Learning classrooms (see Table 6 and Figure 1). We did not find significant ITT effects for WMG (see Table 7) or R2C (see Table 8). When we examined the effect of A2i for kindergarten and first grade students on the LW subtest using HLM, we found significant effects of treatment (p = .004) with an effect size (d) of .37 (see Table 9). We also found a significant effect of grade—first graders had higher scores than kindergarteners. However, there was no grade by treatment interaction effect. When we added EL classroom to the model at the classroom level (see Table 9 bottom), there was still an effect of treatment for A2i with no significant difference for ELs. Nor was there an EL classroom by treatment interaction effect. That is, A2i was effective for improving letter-word recognition of kindergarten and first graders regardless of whether students were in EL or general education classrooms.

When we examined treatment effects for second and third graders on the GM total score, there was no significant effect of treatment (see Table 10). There was a grade effect with third graders achieving generally higher scores than second graders. There was no grade by treatment interaction effect. When we added the EL classroom variable at the classroom level, students in EL classrooms had generally lower GM scores compared with students in general education classrooms, and although there was no significant intent-to-treat main effect, the treatment by EL classroom interaction effect was marginally significant (p = .07). This suggests that EL students experienced larger impacts of A2i on their GM reading scores in second and third grades.

When we added fall pretest scores to the LW and GM impact models and included an interaction between baseline literacy scores and the A2i treatment (see Table 11), we found the following. For LW, the main effect of A2i remained positive and significant (p = .003), and there was no significant interaction with fall LW scores (p = .42). Thus, the impact of A2i was not significantly different for students with higher or lower baseline literacy scores in kindergarten and first grade. For GM, the main effect of A2i was not significant (p = .36), but there was a significant negative coefficient of the interaction with fall GM scores (p = .015). This suggests that the impact of A2i was greater for students with lower initial GM scores.

To summarize, there were not significant ITT effects on the integrated A2i assessments aside from a marginal effect on L2M in kindergarten. However, there was a significant A2i treatment by EL classroom interaction effect on L2M, suggesting that students in EL classrooms benefited more from A2i use than their peers in general education classrooms. There was a significant ITT effect on LW (d = .37; kindergarten and first grade students) scores and no interaction

Table 3
Correlations Among A2i Assessments and Standardized Reading Assessments

Assessment	Spring L2M	Spring WMG	Spring R2C	Fall GM	Spring GM	Fall LW	Spring LW
Spring L2M							
Pearson correlation	1						
N	580						
Spring WMG							
Pearson correlation	.557**	1					
N	561	659					
Spring R2C							
Pearson correlation	.296**	.062	1				
N	283	354	357				
Fall GM							
Pearson correlation	.727**	.370**	.249**	1			
N	249	298	292	304			
Spring GM							
Pearson correlation	.762**	.374**	.355**	.858**	1		
N	256	305	299	285	310		
Fall LW							
Pearson correlation	.645**	.270**	.[a]	.[a]	.[a]	1	
N	274	274	3	0	0	365	
Spring LW							
Pearson correlation	.751**	.294	.[a]	.[a]	.[a]	.859**	1
N	280	279	3	0	0	326	342

Note. L2M = Letters2Meaning assessment; R2C = Reading2Comprehension; GM = Gates MacGinitie Reading Test; LW = Woodcock Johnson III test of Achievements Letter-Word Identification subtest. Grade equivalent scores were used for L2M and R2C. Extended Scale Scores were used for GM and W scores were used for LW.
[a] Cannot be computed because there were no students who took both assessments (LW were administered to K-first graders, GM was administered to second-third graders).
** Correlation is significant at the .01 level (2-tailed).

with EL status. There was not a significant ITT effect of A2i on GM. However, the A2i treatment by students' baseline literacy skills interaction term was significant as was the A2i treatment by EL classroom interaction, suggesting that EL and monolingual students who started the year with less developed reading skills benefited more from the intervention than those who started off as stronger readers.

RQ 3: Relationships Between Teachers' A2i Use and Student Outcomes

We accessed A2i teacher-use logs, which were embedded in the technology to record overall A2i usage, including the time spent using the planning-specific aspects of A2i (i.e., the Literacy Minutes Manager, Student test Scores and the Activity Planner). The user logs serve as a proximal measure of the time individual teachers spent planning for individualized literacy instruction (Connor et al., 2010). It is important to note that this measure cannot provide detailed information about the extent to which teachers adhered to the key recommendations from A2i, only the extent to which they engaged with the technology. However, previous studies using A2i have demonstrated that teacher usage of A2i alongside the fidelity measure of the individualizing student instruction framework is linked with student literacy achievement (Connor, Morrison, Underwood, 2007; Connor et al., 2009).

Considering only the teachers in the immediate treatment condition, we examined whether teachers' time spent using A2i (min) predicted students' spring L2M scores (before the delayed treatment teachers used A2i), controlling for fall L2M scores (see Table 12). We found that the more teachers used A2i over the school year, the greater were their students' word reading skill gains. For every 100 extra minutes teachers spent using A2i, their students' scores generally increased by .1 GEs or about a one-month increase. Importantly, this effect was greater for students with less developed fall scores. Teachers' time spent using A2i had the same effect regardless of whether they were teaching an EL class or a general education class. The fall L2M by EL classroom interaction effect was not significantly different from zero. However, we did not see the same association with two other A2i assessments—WMG and R2C. Furthermore, when these models were run using LW and GM scores, the associations between A2i use and student outcomes were not statistically significant. As such, we do not include detailed tables of model estimates for these two outcomes, but these are available by request to the corresponding author. It is also important to note that because teachers' use of A2i is likely confounded with other factors, the strength of causal inference is considerably weaker than that for the ITT effects.

RQ 3: Examining Teacher Uptake of Professional Development and A2i Use

We next examined teachers' uptake of our PD protocol using the researcher-developed rubric completed. We found that, on average, teachers in the immediate treatment group achieved scores of 30.94 (of 40), which is significantly greater than the teachers in the delayed control condition, who received scores of 22.76 on average. In the immediate treatment condition, teachers in kindergarten and second grade participated in PD more than did teachers in the other grades (kindergarten M = 32.5, SD = 4.1; first grade M = 26.0, SD = 6.6; second grade M = 37.0, SD = 3.6; third grade M = 28.25, SD = 6.9). There was no significant mean difference in teachers' uptake of our PD between EL classrooms and general education classrooms (EL classroom PD uptake M = 28.00, SD = 6.38; general education M = 31.92, SD = 7.09).

Table 4
Baseline Comparisons: Descriptive Statistics and HLM for Kindergarten and First Grade and Second and Third Grade

Condition	*M*	*SD*	*N*
K-1 Fall WJIII Letter-Word Identification			
Delayed treatment	367.46	44.737	165
Immediate treatment	369.51	51.128	162
Total	368.47	47.946	327
2-3 Fall Gates MacGinitie Reading			
Delayed treatment	409.15	36.289	221
Immediate treatment	404.64	38.811	198
Total	407.02	37.525	419

Effect	Estimate	*SE*	95% CI		*p*
			LL	UL	
K-1 Fall WJIII Letter-Word Identification					
Fixed effects					
Intercept	446.637	4.452	437.911	455.363	<.001
A2i immediate treatment	3.095	3.613	-3.986	10.176	.406
Grade	75.634	3.619	68.541	82.727	<.001
EL Class	−12.963	8.630	-29.878	3.952	.134
EL × A2i Immediate Treatment	9.960	7.157	-4.068	23.988	.165
EL × Grade	1.407	7.165	-12.636	15.450	.844

Random effects	*SD*	Variance component	χ^2	*p*
Classroom	5.680	32.262	19.112	.160
Student	29.684	881.158		

Effect	Estimate	*SE*	95% CI		*p*
			LL	UL	
2-3 Fall Gates MacGinitie Reading					
Fixed effects					
Intercept	392.878	6.794	379.562	406.194	<.001
A2i immediate treatment	−6.261	5.205	−16.463	3.941	.249
Grade	26.934	5.241	16.662	37.206	<.001
EL Class	−27.110	12.442	−51.496	−2.724	.030
EL × A2i Immediate Treatment	18.996	10.177	−0.951	38.942	.063
EL × Grade	4.329	10.195	−15.653	24.311	.671

Random effects	*SD*	Variance component	χ^2	*p*
Classroom	0.597	0.356	10.410	>.500
Student	29.474	868.742		

Note. Kindergarten and first grade descriptive statistics for letter-word identification at baseline. Wilks' Lambda = .999, p = .844. W-scores were used. (First Table). Second and third grade descriptive statistics for Gates MacGinitie Reading Test at baseline. $F(1, 417) = 1.51$, p = .219. Extended Scaled Scores were used (Second Table). HLM for kindergarten and first grade fall letter-word identification and for second and third grade fall Gates-MacGinitie Reading Test (Third Table). A2i immediate treatment = 1; Delayed Treatment = 0. Grade coded K = 0, 1st grade = 1 or 2nd grade = 0 and 3rd grade = 1. WJII Model Deviance = 3471.262866. GM Model Deviance = 2895.465514.

Overall, the teachers in the immediate treatment condition (n = 16) used A2i for an average of 161.30 minutes (SD = 84.62), and this ranged from a low of 64 minutes to a high of 425 minutes. We did not compare this with the teachers in the delayed treatment condition because they only had access to A2i from April through June. There was no significant difference in the amount of time teachers spent using A2i when we compared EL classrooms with general education classrooms. General education classroom teachers used A2i for a mean of 171.3 minutes (95% CI [118.28, 224.30]), whereas EL classroom teachers used A2i for a mean of 131.33 (95% CI [39.51, 223.14]) minutes. Finally, we found a significant correlation between PD uptake and use of A2i (min, r = .526, p = .037). That is, the more teachers participated in PD, the more likely they were to spend time using A2i, or similarly, teachers who used A2i were more likely to participate in PD.

Discussion

The purpose of this study was to describe the process of bringing A2i to scale for effective implementation in classrooms serving a linguistically diverse group of learners. We investigated teachers' use of the revised A2i technology with PD support on the literacy outcomes of English monolingual students

Table 5

Intent-to-Treat Effects for Letters2Meaning (L2M) in Kindergarten–Third Grade: HLM Model Predicting Treatment Effects on Spring L2M Scores

Effect	Estimate	*SE*	95% CI LL	95% CI UL	*p*
		L2M Spring scores			
Fixed effects					
Intercept	0.863	0.039	0.787	0.939	<.001
A2i immediate treatment	0.097	0.056	−0.013	0.207	.090
Grade	0.648	0.031	0.587	0.709	<.001
Grade × A2i Immediate Treatment	−0.033	0.042	−0.115	0.049	.434

Random effects	*SD*	Variance component	χ^2	*p*
Classroom	0.03,719	0.001	31.779	.428
Student	0.64,074	0.411		

Note. A2i = Assessment-to-Instruction. A2i immediate treatment = 1; Delayed Treatment = 0. Grade coded K = 0, first grade = 1, second grade = 2 and third grade = 3. Model Deviance = 1,138.935716.

and ELs. We contextualized our process of reaching effective scalability by grounding the study in the EPIS model, a conceptual framework within the Implementation Science field. We present our findings in one article to illustrate a path to effective classroom change, spanning from redesigning to implementing A2i. Our findings are important, because they provide a theoretical framework, with specific practices and procedures, for researchers and practitioners to implement evidence-based interventions within classrooms. These data also provide initial evidence of consequential validity of A2i, such that, documenting teachers' use of A2i is what makes the tool scalable and leads to meaningful change when used as intended within classrooms. Overall, our newly designed A2i technology, including the new DFI algorithms, shows promise to use at scale with kindergarten and first grade monolingual students and ELs. We have five principle findings gleaned from the two phases of the study:

(1) Our aim was to develop computer-based adaptive assessments that teachers could administer easily and that were valid and reliable for linguistically diverse students. In general, results show that the integrated A2i online adaptive assessments were psychometrically strong; particularly WMG and L2M (see Appendix C). R2C had limited range and was appropriate only for students with strong reading skills. This result suggests the need to develop more R2C items to assess students with varying reading abilities. Currently, only students in second or third grade are able to take the R2C assessment. Furthermore, we found that teachers required more support from the research team to use the assessments independently than anticipated, but there was variability with some teachers able to use the assessments independently while others not at all. One reason for this may have been due to the content within our PD session. We primarily focused on helping teachers read and interpret assessment results to plan individualized instruction for their students, with little focus on the logistics of administering the assessments. With our goal of sustainability, we plan to include more PD centered on assessment administration (i.e., logging into A2i, navigating the assessments) to ensure that

Table 6

HLM Model Predicting ITT Effects on Spring L2M Scores in Kindergarten–Third Grade, Including EL as a Classroom Level Variable

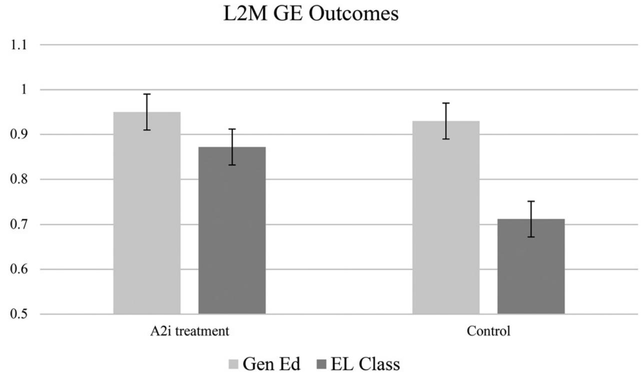
Effect	Estimate	*SE*	95% CI LL	95% CI UL	*p*
		L2M Spring scores			
Fixed effects					
Intercept	0.930	0.044	0.844	1.016	<.001
A2i immediate treatment	0.021	0.052	−0.081	0.123	.695
Grade	0.637	0.020	0.598	0.676	<.001
EL Class	−0.218	0.053	−0.322	−0.114	<.001
EL × A2i Immediate Treatment	0.142	0.069	0.007	0.277	.048

Random effects	*SD*	Variance component	χ^2	*p*
Classroom	0.009	<0.001	25.204	>.500
Student	0.637	0.406		

Note. L2M = Letters to Meaning; A2i = Assessment-to-Instruction; HLM = hierarchical linear model; EL = English learners; ITT = intent-to-treat. A2i immediate treatment = 1; Delayed Treatment = 0. Grade coded K = 0, first grade = 1, second grade = 2 and third grade = 3. Model Deviance = 1,137.247867.

Figure 1

Modeled Results for Kindergarten in General Education (General Education Classroom in Light Gray) and English Immersion Classrooms (EL Classroom in Dark Gray)



Note. Results look the same for first grade, but scores are higher. Error bars are standard errors. L2M GE = Letters2Meaning grade equivalent scores.

teachers have the foundational knowledge they need to move forward independently.

In addition, some teachers questioned the validity of the newly developed assessments and the results, feeling that their students' scores were too low overall. Fortunately, the IRT results demonstrate that the integrated online assessments are valid and reliable and correlate significantly with the LW and GM assessments, which are widely used standardized measures of reading. Yet, this example highlights the importance of the "preparation" stage in the EPIS model; making sure that the climate is ready for implementation includes fostering trust and "buy-in" from teachers. If teachers question the validity of an intervention, for example, they will likely not believe that implementing the tool will benefit themselves or their students. The teachers' view that the assessments underestimated their students' abilities, therefore, points to the need to better prepare teachers in observing, understanding, interpreting variability in their students' individual skill development, such as stronger word decoding skills than vocabulary skills as we observed in our sample.

(2) Overall, we observed mixed results for the quasi-experimental intent-to-treat analyses. The standardized reading assessments (LW for kindergarten and first grade; GM for second and third grade) revealed that the intent-to-treat effect was significant only for kindergarten and first grade students (d = .15 L2M, p = .09*;* d = .37 LW, p = .004). There was no main treatment effect for second and third graders on the GM; however, there is evidence of interactions between A2i intervention with EL status and baseline literacy scores. Students in EL classrooms and those with less developed literacy skills in second and third grade experienced larger gains when their teachers used A2i.

We present two possible interpretations of the differential findings we observed in kindergarten and first grade compared with second and third grade. The first possibility is that an "active ingredient" of A2i implementation may be appropriately timed individualized code-focused instruction. Because the development of code-focused skills is critical during kindergarten and first grade, it is possible that better alignment between students' instructional needs with the actual instruction they are provided leads to better overall word reading outcomes. Intervention that primarily affects code-focused skills may be less effective for students in subsequent elementary grades who are starting to make the transition from *learning to read* to *reading to learn* (Chall, 1996; Wanzek et al., 2010)—those who may be nearing mastery of code-focused skills. This interpretation is supported by previous studies that have documented the effects of A2i on code-focused skills (e.g., Al Otaiba et al., 2011; Connor, Morrison, Underwood, 2007; Connor et al., 2007; Connor et al., 2013; Connor, Morrison, Fishman, et al., 2011; Connor, Morrison, Schatschneider, et al., 2011) as well as studies that have documented A2i treatment effects on reading comprehension outcomes in third grade (Connor, Morrison, Fishman, et al., 2011). Furthermore, in a longitudinal efficacy study evaluating A2i, Connor and colleagues (2013) found that effects may be cumulative such that, unless second and third graders had participated in A2i classrooms beginning in first grade, their performance was not significantly different than students in the control condition

Table 7

Intent-to-Treat Effects for Word Match Game (WMG) in Kindergarten–Third Grade: HLM Model Predicting Treatment Effects on Spring WMG Scores

			95% CI		
Effect	Estimate	*SE*	LL	UL	*p*
	WMG Spring scores				
Fixed effects					
Intercept	0.291	0.051	0.191	0.392	<.001
A2i immediate treatment	0.006	0.065	−0.122	0.134	.928
Grade	0.390	0.042	0.308	0.473	<.001
Grade × A2i Immediate Treatment	0.019	0.050	−0.080	0.118	.710

Random effects	*SD*	Variance component	χ^2	*p*
Classroom	0.136	0.018	51.209	.013
Student	0.679	0.461		

Note. A2i = Assessment-to-Instruction; HLM = hierarchical linear model. A2i immediate treatment = 1; Delayed Treatment = 0. Grade coded K = 0, first grade = 1, second grade = 2 and third grade = 3 and grand mean centered. Model Deviance = 1,390.535888.

Table 8

Intent-to-Treat Effects for Reading2Comprehension (R2C) in Second and Third Grade: HLM Model Predicting Treatment Effects on Spring R2C Scores

			95% CI		
Effect	Estimate	*SE*	LL	UL	*p*
R2C Spring scores					
Fixed effects					
Intercept	1.249	0.073	1.105	1.393	<.001
A2i immediate treatment	−0.033	0.087	−0.203	0.136	.700
Grade	0.128	0.101	−0.069	0.325	.220
Grade × A2i Immediate Treatment	0.077	0.118	−0.156	0.309	.518

Random effects	*SD*	Variance component	χ^2	*p*
Classroom	0.087	0.008	17.509	>.500
Student	0.671	0.451		

Note. A2i = Assessment-to-Instruction; HLM = hierarchical linear model. A2i immediate treatment = 1; Delayed Treatment = 0. Grade coded second grade = 2 and third grade = 3 and grand mean centered. Model Deviance = 740.505185.

(teachers did not use A2i). Thus, as previous studies have indicated, a clear recommendation from an implementation standpoint is to introduce A2i starting in kindergarten and first grade and then follow students into second and third grade in a gradual rollout. Such gradual rollout could also support sustainment, the final step of the EPIS model (Aarons et al., 2011).

A second possible explanation for the differential findings by grade lies in the outcome measure for determining intent-to-treat effects for students in the lower vs. upper grades. In kindergarten and first grades, the intent-to-treat effect was based on students' word reading, whereas in second and third grade, a standardized measure of reading comprehension was used. Word reading is a relatively more constrained skill set that is more malleable with effective instruction, at least in the short-term, than reading comprehension (e.g., Paris, 2005; Snow & Matthews, 2016). Word reading is also easier to measure relative to reading comprehension, because the scope and sequence of development are more clearly defined (Snow & Matthews, 2016) and there is less susceptibility to bias since word reading is significantly less dependent on factors like background knowledge or inference making skills (e.g., Kim, 2017, 2020). In contrast, reading comprehension is a notoriously difficult construct to change and to measure. Even the most rigorous of studies, such as those conducted through the Reading for Understanding initiative, found it difficult to "move the needle" on reading comprehension (Pearson et al., 2020). Nevertheless, prior research on A2i did find significant impacts on reading comprehension (Connor, Morrison, Underwood, 2007; Connor, Morrison, Fishman, et al., 2011). Considering the small sample of the present study, its quasi-experimental design, and the relatively short duration of teachers' implementation of A2i, it is not surprising that we did not observe a significant treatment effect with reading comprehension as the outcome variable.

We have certainly considered what modifications would need to be made to A2i to bring about a strong treatment effect when comprehension is the outcome variable. We concede that quantity (i.e., recommendations in minutes of time spent in a given instructional activity) is only one element of instructional quality. It is likely that, to make significant changes in children's comprehension abilities, a coherent, knowledge-rich curriculum that builds knowledge within and across grade levels will be necessary (Hirsch, 2006; Kamhi & Catts, 2017; Willingham, 2006). In addition, we face the same barriers as other researchers in this area in finding ways to properly assess reading comprehension by either (a) aligning reading comprehension assessment closely to actual content being taught or (b) finding ways to decouple background knowledge from reading comprehension performance (to the extent possible) such as with a more authentic assessment like the GISA (O'Reilly, Sabatini, & Deane, 2013). Although there is clearly more work to be done to improve reading comprehension, we found it promising that students with less developed reading abilities benefited from participating in classrooms where A2i was being used (the A2i treatment by students' baseline literacy skills interaction effect as measured by GM performance). Nonetheless, these findings highlight the need for a continued focus on improving effective instructional practices for promoting growth of meaning-focused skills, which might be particularly important for ELs since limited L2 oral language proficiency may interfere with successful L2 reading comprehension (e.g., Lesaux, 2006; Mancilla-Martinez & Lesaux, 2010; Nakamoto et al., 2007).

(3) We found promising effects for ELs. Our findings provide convincing evidence that A2i usage leads to improved word reading outcomes for ELs, with greater effects for students in EL classrooms than for students in general education classrooms. This finding is supported in the research literature. Studies have found that although ELs often enter school with less developed literacy skills (Hammer et al., 2011), they are able to perform on par with their monolingual peers on word-reading accuracy after as little as one year of formal instruction (see Lesaux & Geva, 2006, for a review). In addition, we documented a marginally significant EL by A2i treatment interaction effect on students' reading comprehension outcomes. This interaction effect suggests that meaning-focused, individualized instruction may also lead to improved reading comprehension outcomes. This finding supports a central theme in the literature on effective literacy instruction for ELs—namely, that instruction focusing on the development of oral language skills is integral for successful reading comprehension (August & Shanahan, 2006; Castro et al., 2011). Hence, these findings provide

Table 9

Spring WJ Letter-Word Identification Descriptive Statistics (Top) and HLM Intent-to-Treat Effect for Kindergarten and First Grade (Middle) and Adding EL Classroom (Bottom)

Descriptive statistics	Grade	*M*	*SD*	*N*
Delayed treatment	Kindergarten	383.22	24.774	88
	First	428.46	27.519	87
	Total	405.71	34.581	175
A2i immediate treatment	Kindergarten	395.89	25.593	88
	First	436.51	30.047	76
	Total	414.71	34.321	164
Total	Kindergarten	389.55	25.906	176
	First	432.21	28.918	163
	Total	410.06	34.699	339

Effect	Estimate	*SE*	95% CI LL	95% CI UL	*p*
HLM Results					
Fixed effects					
Intercept	384.650	2.517	379.717	389.583	<.001
A2i Immediate Treatment	10.090	2.935	4.337	15.843	.004
Grade	42.362	2.936	36.607	48.117	<.001

Random effects	*SD*	Variance component	χ^2	*p*
Classroom	0.427	0.182	9.228	>.500
Student	27.045	731.433		

Effect	Estimate	*SE*	95% CI LL	95% CI UL	*p*
HLM Results Including EL Interaction					
Fixed effects					
Intercept	386.089	2.733	380.732	391.446	<.001
A2i immediate treatment	7.706	3.393	1.056	14.356	.044
Grade	42.741	2.951	36.957	48.525	<.001
EL class	−6.944	4.843	−16.436	2.548	.179
EL × A2i Immediate Treatment	9.797	6.799	−3.529	23.123	.177

Random effects	*SD*	Variance component	χ^2	*p*
Classroom	0.429	0.184	6.820	>.500
Student	27.029	730.552		

Note. WJ = The Woodcock-Johnson III Test; A2i = Assessment-to-Instruction; HLM = hierarchical linear model; EL = English learners. W-scores were used. A2i immediate treatment = 1; Delayed Treatment = 0. Grade coded K = 0, first grade = 1. Model Deviance = 3,198.329921.

preliminary evidence suggesting that both code- and meaning focused instruction, when aligned with ELs' individualized needs, may lead to improved literacy skills, and that A2i can leverage knowledge of ELs' baseline skills to lead to better individualized instruction. This discussion must be tempered with the caution that there was a single EL classroom per grade level so differential effects would have to be quite large to detect them within this study design. Taken together, the revised DFI algorithms used in A2i appear to be working as expected in kindergarten and first grade classrooms. Hence, using A2i technology to individualize student instruction shows promise of efficacy for both English monolingual students and ELs.

(4) Analyses of the relationship between students' postintervention reading scores and teachers' time spent with A2i technology revealed that the more teachers used A2i (min), the greater were their kindergarten through third grade students' reading gains on one A2i assessment of word reading. This relationship was stronger for students with less developed reading skills in the fall. Teachers in EL classrooms generally used A2i to the same extent as teachers in general education classrooms. Moreover, this effect of A2i use within the A2i immediate treatment condition was consistent for students in EL classrooms and in general education classrooms. However, these findings were not replicated when measures of vocabulary or reading comprehension were used as the outcome measure. These results suggest that A2i technology can be used in classrooms that serve students from diverse linguistic backgrounds with varying levels of English proficiency to improve word reading, and the revised DFI algorithms are working as anticipated. The data also suggest that teachers in both general education and EL classrooms are able to better support student development of code-focused skills through individualizing instruction using A2i.

Table 10

Spring Gates MacGinitie Reading Test Descriptive Statistics (Top) and HLM Intent-to-Treat Effect for Second and Third Grades (Middle) and Adding EL Classroom (Bottom)

Descriptive statistics	Grade	*M*	*SD*	*N*
Delayed treatment	Second	418.92	41.252	61
	Third	451.39	40.937	83
	Total	437.63	43.979	144
A2i immediate treatment	Second	412.06	35.006	83
	Third	444.22	27.163	83
	Total	428.14	35.153	166
Total	Second	414.97	37.792	144
	Third	447.80	34.820	166
	Total	432.55	39.717	310

			95% CI		
Effect	Estimate	*SE*	LL	UL	*p*
HLM Results					
Fixed effects					
Intercept	405.843	8.924	388.352	423.334	<.001
A2i immediate treatment	−8.651	7.127	−22.620	5.318	.245
Grade	31.663	7.147	17.655	45.671	<.001

Random effects	*SD*	Variance component	χ^2	*p*
Classroom	11.979	143.505	42.279	<.001
Student	34.603	1,197.395		

			95% CI		
Effect	Estimate	*SE*	LL	UL	*p*
HLM Results Including EL Interaction					
Fixed effects					
Intercept	414.000	7.281	399.729	428.271	<.001
A2i immediate treatment	−14.246	6.224	−26.445	−2.047	.041
Grade	30.490	5.388	19.930	41.050	<.001
EL class	−29.179	8.725	−46.280	−12.078	.006
EL × A2i Immediate Treatment	24.010	12.245	0.010	48.010	.074

Random effects	*SD*	Variance component	χ^2	*p*
Classroom	7.211	52.004	21.630	.042
Student	34.638	1,199.806		

Note. A2i = Assessment-to-Instruction; HLM = hierarchical linear model; EL = English learners. A2i immediate treatment = 1; Delayed Treatment = 0. Grade was grand mean centered with grade 2 = 2 and grade 3 = 3. EL classroom = 1; General Education Classroom = 0. Deviance = 3,053.991523.

(5) Overall, teachers' uptake of our PD varied by grade level, with kindergarten and second grade teachers more likely to participate in the A2i PD protocol than first and third grade teachers. There is not a theoretical reason we can ascertain that would explain this grade level effect based on the nature of teaching other than individual variation. Important aspects of practitioner level variables from the EPIS model, such as the openness to change, the conviction that change needs to happen in order for goals to be met, and different perceptions of risk to change could explain these grade level differences (Aarons et al., 2011). Furthermore, uptake did not vary between EL classrooms and general education classrooms. The more teachers, including teachers of EL classrooms, participated in the A2i PD, the more likely they were to spend using A2i. Furthermore, as noted above, the more time teachers used A2i, the greater were their students' word reading skill gains. However, we did not find a similar relationship with other measures of reading (e.g., vocabulary, reading comprehension).

In scaling up the PD, we attempted to move resources online so they were easily accessible through A2i. However, the PD was still too expensive to be fully scalable. Cost analyses suggested that with the current PD protocol, the entire implementation cost per student is about $150 (including PD, technical support, administrative support, etc.), noting that PD is the primary driver of implementation costs. A more scalable version would have total implementation costs closer to about $50 per student. It may be that moving more of the PD online and replacing most face-to-face interactions with video conferencing would reduce costs while maintaining efficacy. In light of the COVID-19 pandemic, accessibility of effective web-based interventions, such as A2i, for individualizing student learning is increasingly important. Although our initial concerns with

Table 11

ITT Effects Including Baseline Literacy Interaction for Spring WJ Letter-Word Identification in Kindergarten and First Grades (Top) and Spring Gates MacGinitie Reading Test in Second and Third Grades (Bottom)

			95% CI		
Effect	Estimate	*SE*	LL	UL	*p*
K-1 Spring WJIII Letter-Word Identification					
Fixed effects					
Intercept	385.068	2.235	380.687	389.449	<.001
A2i immediate treatment	9.570	2.570	4.533	14.607	.003
Grade	41.991	2.569	36.956	47.026	<.001
Fall LW	0.781	0.048	0.687	0.875	<.001
Fall LW × A2i Immediate Treatment	−0.051	0.063	−0.174	0.072	.422

Random effects	*SD*	Variance component	χ^2	*p*
Classroom	3.644	13.278	26.197	.016
Student	16.235	263.586		

			95% CI		
Effect	Estimate	*SE*	LL	UL	*p*
Grd. 2–3 Spring Gates MacGinitie Reading					
Fixed effects					
Intercept	359.097	20.035	319.828	398.366	<.001
A2i immediate treatment	−7.067	7.552	−21.869	7.735	.365
Grade	30.374	7.560	15.556	45.192	.001
Fall GM	1.053	0.064	0.928	1.178	<.001
Fall GM × A2i Immediate Treatment	−0.202	0.082	−0.363	−0.041	.015

Random effects	*SD*	Variance component	χ^2	*p*
Classroom	14.631	214.066	127.577	<.001
Student	19.792	391.712		

Note. WJ = The Woodcock-Johnson III Test; LW = Letter-Word; A2i = Assessment-to-Instruction; GM = The Gates-MacGinitie Reading Test. A2i immediate treatment = 1; Delayed Treatment = 0. Grade coded K = 0, first grade = 1 or second grade = 2 and third grade = 3 and grand mean centered. WJII model Deviance = 2,745.773536, GM model Deviance = 2,535.290981.

scalability revolved around pricing, schools now face the added challenge of distance learning, often mediated by technology. This necessitates such web-based approaches as A2i to support classroom learning for all students, including linguistically diverse students.

Limitations

There are limitations to this study that should be considered when interpreting the results. We conducted a quasi-experiment where two schools were randomly assigned to immediate or delayed treatment conditions. However, in the analyses, treatment variables were entered at the classroom level rather than at the school level, and the numbers of classrooms and students in this study were small with regards to power for subgroup and moderation analyses. A fully powered randomized, controlled trial was beyond the scope of the project; however, the groups were equivalent at baseline, which is a strength. To examine the efficacy of A2i and its impact on diverse student populations, a fully powered randomized controlled trial is needed. Next, we intentionally recruited higher poverty schools that served a higher proportion of Hispanic/Latinx students, with approximately 25% of students in English immersion classrooms based on their reported limited English proficiency. Unfortunately, the schools would not allow us to assess students' language and reading skills in Spanish, so we relied on the schools' assessment of English proficiency. There were certainly dual language learners (i.e., students from non-English speaking homes) in the general education classrooms, but we were not able to identify them. Thus, we had to rely on school report on students' EL status as a classroom-level variable (i.e., EL classroom). Additionally, it is not clear that these findings would generalize to other school settings with different student demographics and varying levels of teachers' openness to innovation although studies using the research version of A2i suggest that A2i and individualizing student instruction is effective across a range of school settings (e.g., Connor, Morrison, Underwood, 2007; Connor, Morrison, Fishman, et al., 2011; Connor et al., 2013). Furthermore, we acknowledge that the measures we used to assess teachers' uptake of PD and A2i usage could be improved. To fully understand how fidelity of implementation impacts student outcomes, more information about how teachers use A2i's recommendations in the classroom is needed. Simply knowing the amount of time teachers spent using A2i can only give us a measure of surface fidelity and more sophisticated analyses (e.g., mediational or instrumental variable analyses) would be needed to fully understand this relationship. As a future direction, we plan to design and

Table 12
Predicting Effects of A2i Use (Minutes) on Students' Spring L2M Outcomes for Immediate Treatment Group Only (Top) and Considering EL Classrooms (Bottom)

			95% CI		
Effect	Estimate	*SE*	LL	UL	*p*
		HLM Results			
Fixed effects					
Intercept	1.891	0.140	1.616	2.166	<.001
A2i Use	0.004	0.002	0.001	0.008	.023
Fall L2M	0.275	0.041	0.194	0.355	<.001
A2i Use × Fall L2M	−0.001	0.000	−0.002	−0.001	<.001
Random effects	*SD*	Variance component		χ^2	*p*
Classroom	0.536	0.287		207.262	<.001
Student	0.563	0.317			

			95% CI		
Effect	Estimate	*SE*	LL	UL	*p*
		HLM Results Including EL			
Fixed effects					
Intercept	1.856	0.207	1.450	2.262	<.001
A2i Use	0.005	0.002	0.000	0.009	.055
EL Class	0.150	0.422	−0.677	0.976	.728
Fall L2M	0.288	0.050	0.189	0.387	<.001
A2i Use × Fall L2M	−0.001	0.000	−0.002	−0.001	<.001
EL Class × Fall L2M	−0.105	0.100	−0.300	0.090	.292
Random effects	*SD*	Variance component		χ^2	*p*
Classroom	0.694	0.482		331.145	<.001
Student	0.563	0.317			

Note. L2M = Letters to Meaning; A2i = Assessment-to-Instruction; HLM = hierarchical linear model; EL = English learners. W-scores were used. A2i immediate treatment = 1; Delayed Treatment = 0. Grade coded K = 0, first grade = 1, second grade = 2 and third grade = 3. Model Deviance (top) = 443.122457, model Deviance (bottom) = 450.328206.

cross-validate measures of teachers' uptake that would more carefully examine teachers' behaviors in relation to their student outcomes. Finally, we acknowledge that the adjustments made to the A2i algorithms may not reflect the full set of language and literacy needs ELs bring to the classroom; rather, they were developed to use with both monolingual students and ELs. As mentioned above, we were limited by outer context factors of AZ state laws on monolingual instruction and assessment. As an area for future research, we plan to develop a partner set of validated assessments in students' first language (Spanish in this case) that would make recommendations in light of students' first and second language and literacy abilities. Beyond this, it is our goal that the algorithms will provide recommendations for both English and Spanish instruction (i.e., dual language instruction) with an eye toward supporting biliterate readers.

Other Lessons Learned and Scaling Up

Principal buy-in, the extent to which principals supported and enforced the school-wide implementations of A2i for individualizing instruction, was found to be instrumental in ensuring teachers' use of A2i. This lesson is confirmed in the EPIS framework idea of stakeholders who act as inner contextual factors to promote and lead change (Aarons et al., 2011). Grade level teams engaged more with the technology when there was at least one teacher at a grade level team that advocated for the use of A2i. Thus, we strongly recommend that for scale up, implementation be focused on the entire system—district, school, and classroom. This might include memos of understanding with the district and identifying literacy champions at the school to work closely with teachers and literacy coaches. Implementation Science suggests that such a strategy should be effective (Fixsen et al., 2013).

A critical finding of this study was that in kindergarten and first grade, A2i was effective for improving students' word reading skills in EL classrooms and was similarly effective for students in general education classrooms. Moreover, there was no significant difference in outcomes in kindergarten and first grade for EL and monolingual students, which is highly encouraging. According to the Census, ELs now make up 25% of elementary students (Bauman, 2017). Thus, studies that identify potentially effective, scalable interventions must logically include analysis for linguistically diverse students. Based on the proportionality of ELs in classrooms, and their unique needs, educational programs that do not consider ELs are less likely to be successfully implemented at scale.

There is ongoing debate about the "Science of Reading" and how to support teachers' use of evidence-based practices (e.g., Castles et al., 2018; Solari et al., 2020). The "reading wars" were the original inspiration for A2i, and, regrettably, the battle has become reinvigorated. A2i is positioned to answer this reemerging challenge of supporting teachers in providing effective literacy instruction given A2i's long track record of efficacy (e.g., Connor et al., 2004; Connor, Morrison, Fishman et al., 2007), and now, initial implementation research. A concern about the science of reading movement is that there is not clear advice to teachers about exactly what the science of reading looks like in their classrooms. Although data driven, individualized instruction is associated with substantial literacy gains (e.g., Al Otaiba et al., 2011; Fuchs et al., 1994), teachers find it difficult to implement effectively (Roehrig et al., 2008). A2i technology facilitates individualized instruction and supports the delivery of more efficacious and efficient instruction (e.g., Connor et al., 2009; Connor, Morrison, Schatschneider et al., 2011). A compelling reason to get effective interventions, such as A2i, off of researchers' computers and into classrooms is to provide effective tools for teachers that operationalize the science of reading in ways that ensure that students achieve proficient reading skills.

This study describes the process of bringing A2i technology to scale using the EPIS implementation model. We outline the lessons we learned, providing a framework for future research and practice. With funding through the Department of Education, Education Innovation Research (EIR) program, we are currently using these data to plan and conduct a large-scale study to bring A2i to scale nationally at a reasonable cost per student. This means that A2i could potentially move from being a pure research tool to a professional support system that can be used in many schools that differ substantially in location (e.g., New Jersey, New York, Pennsylvania, California) and student populations (although the focus of the EIR project is on working with schools that serve children in need). In the EIR project, we added an out-of-school component so that individualized student learning experiences could continue in students' homes and communities (learningovations.com; readcharlotte.org). This focus has become more critical during the COVID-19 pandemic, because much of the instruction and learning experiences happen in out-of-school contexts and online domains. The results of the current scalability study described throughout this article directly inform what we are now doing in the EIR project nationally. It is our intention that these studies, together, provide an example of the EPIS model in school settings for other researchers and practitioners as they work to bring their effective programs to scale.

Tribute to Dr. Carol Connor

Dr. Carol Connor passed away May 14, 2020, while revising the final version of this article. True to her diligent nature, she worked until the day before she passed away after battling cancer. Implementing Assessment-to-Instruction (A2i) in schools across the nation was Dr. Connor's dream, as her life's work was centered on ensuring that all students read proficiently by the 3rd grade. She developed the Individualized Student Instruction (ISI) framework and patented the algorithms that drive A2i, publishing her first of many randomized controlled trials in 2005 to support this work. Dr. Connor has made extraordinary impact both nationally and internationally. She truly touched the lives of everyone who had the honor to cross her path. Her legacy will live on through the mentees and colleagues who have had the privilege of knowing and working with her and through A2i, as educators across the nation use the technology to individualize instruction for their students. Most importantly, her legacy will live on through the many children who have and will continue to benefit from ISI and A2i. Dr. Carol Connor did not work hard for accolades. This became increasingly apparent as her time was running out. What she worked for was the future of others. She will be deeply missed everyday but never forgotten.

References

Aarons, G. A., Hurlburt, M., & Horwitz, S. M. (2011). Advancing a conceptual model of evidence-based practice implementation in public service sectors. *Administration and Policy in Mental Health*, *38*(1), 4–23. https://doi.org/10.1007/s10488-010-0327-7

Al Otaiba, S., Connor, C. M., Folsom, J. S., Greulich, L., Meadows, J., & Li, Z. (2011). Assessment data-informed guidance to individualize kindergarten reading instruction: Findings from a cluster-randomized control field trial. *The Elementary School Journal*, *111*(4), 535–560. https://doi.org/10.1086/659031

Albro, E. R. (2020). *How IES advances education research*. [Paper presentation]. Department of Education, Institute of Education Sciences, Washington DC. https://www.aera.net/Newsroom/AERA-Highlights-E-newsletter/-em-AERA-Highlights-em-October-2017/Q-A-IESs-Liz-Albro-Discusses-How-IES-Advances-Education-Research

August, D., Artzi, L., & Barr, C. (2016). Helping ELLs meet standards in English language arts and science: An intervention focused on academic vocabulary. *Reading & Writing Quarterly*, *32*(4), 373–396. https://doi.org/10.1080/10573569.2015.1039738

August, D., Artzi, L., Barr, C., & Francis, D. (2018). The moderating influence of instructional intensity and word type on the acquisition of academic vocabulary in young English language learners. *Reading and Writing*, *31*(4), 965–989. https://doi.org/10.1007/s11145-018-9821-1

August, D., McCardle, P., & Shanahan, T. (2014). Developing literacy in English language learners: Findings from a review of the experimental research. *School Psychology Review*, *43*(4), 490–498. https://doi.org/10.1080/02796015.2014.12087417

August, D., & Shanahan, T. (2006). *Developing literacy in second-language learners: Report of the national literacy panel on language minority children and youth*. Lawrence Erlbaum Associates.

August, D., & Shanahan, T. (2010). Response to a review and update on developing literacy in second-language learners: Report of the national literacy panel on language minority children and youth. *Journal of Literacy Research*, *42*(3), 341–348. https://doi.org/10.1080/1086296X.2010.503745

Baker, D. L., Basaraba, D. L., & Polanco, P. (2016). Connecting the present to the past: Furthering the research on bilingual education and bilingualism. *Review of Research in Education*, *40*(1), 821–883. https://doi.org/10.3102/0091732X16660691

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

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

Bialystok, E. (2001). *Bilingualism in development: Language, literacy, and cognition*. Cambridge University Press. https://doi.org/10.1017/CBO9780511605963

Bos, C., Mather, N., Narr, R. F., & Babur, N. (1999). Interactive, collaborative professional development in early literacy instruction: Supporting the balancing act. *Learning Disabilities Research & Practice*, *14*(4), 227–238. https://doi.org/10.1207/sldrp1404_4

Calderon, M., Hertz-Lazarowitz, R., & Slavin, R. E. (1998). Effects of bilingual cooperative integrated reading and composition on students making the transition from Spanish to English reading. *The Elementary School Journal*, *99*(2), 153–165. https://doi.org/10.1086/461920

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5–51. https://doi.org/10.1177/1529100618772271

Castro, D. C., Páez, M. M., Dickinson, D. K., & Frede, E. (2011). Promoting language and literacy in young dual language learners: Research, practice, and policy. *Child Development Perspectives*, *5*(1), 15–21. https://doi.org/10.1111/j.1750-8606.2010.00142.x

Chall, J. S. (1996). *Stages of reading development* (2nd ed.). Harcourt Brace.

Cheung, A. C. K., & Slavin, R. E. (2012). Effective reading programs for Spanish-dominant English language learners (ELLs) in the elementary grades: A synthesis of research. *Review of Educational Research*, *82*(4), 351–395. https://doi.org/10.3102/0034654312465472

Collins, B. A. (2014). Dual language development of Latino children: Effect of instructional program type and the home and school language environment. *Early Childhood Research Quarterly*, *29*(3), 389–397. https://doi.org/10.1016/j.ecresq.2014.04.009

Common Core State Standards Initiative. (2010). *Common Core State Standards for English language arts & literacy in history/social studies, science, and technical subjects*. Pennsylvania Department of Education.

Connor, C. M. (2019). Using Technology and Assessment to Personalize Instruction: Preventing Reading Problems. *Prevention Science: The official Journal of the Society for Prevention Research*, *20*(1), 89–99. https://doi.org/10.1007/s11121-017-0842-9

Connor, C. M. (2016). A lattice model of the development of reading comprehension. *Child Development Perspectives*, *10*(4), 269–274. https://doi.org/10.1111/cdep.12200

Connor, C. M., & Morrison, F. J. (2016). Individualizing student instruction in reading: Implications for policy and practice. *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 54–61. https://doi.org/10.1177/2372732215624931

Connor, C. M., Day, S. L., Phillips, B., Sparapani, N., Ingebrand, S. W., McLean, L., Barrus, A., & Kaschak, M. P. (2016). Reciprocal effects of self-regulation, semantic knowledge, and reading comprehension in early elementary school. *Child Development*, *87*(6), 1813–1824. https://doi.org/10.1111/cdev.12570

Connor, C. M., Morrison, F. J., & Katch, L. E. (2004). Beyond the reading wars: Exploring the effect of child × instruction interactions on growth in early reading. *Scientific Studies of Reading*, *8*(4), 305–336. https://doi.org/10.1207/s1532799xssr0804_1

Connor, C. M., Morrison, F. J., & Underwood, P. (2007). A second chance in second grade? The independent and cumulative Impact of first and second grade reading Instruction and students' letter-word reading skill growth. *Scientific Studies of Reading*, *11*(3), 199–233. https://doi.org/10.1080/10888430701344314

Connor, C. M., Morrison, F. J., Fishman, B. J., Schatschneider, C., & Underwood, P. (2007). The early years. Algorithm-guided individualized reading instruction. *Science*, *315*(5811), 464–465. https://doi.org/10.1126/science.1134513

Connor, C. M., Morrison, F. J., Fishman, B., Crowe, E. C., Al Otaiba, S., & Schatschneider, C. (2013). A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade. *Psychological Science*, *24*(8), 1408–1419. https://doi.org/10.1177/0956797612472204

Connor, C. M., Morrison, F. J., Fishman, B., Giuliani, S., Luck, M., Underwood, P., & Schatschneider, C. (2011). Classroom instruction, child × instruction interactions and the impact of differentiating student instruction on third graders' reading comprehension. *Reading Research Quarterly*, *46*(3), 189–221. https://doi.org/10.1598/RRQ.46.3.1/epdf

Connor, C. M., Morrison, F. J., Schatschneider, C., Toste, J., Lundblom, E., Crowe, E. C., & Fishman, B. (2011). Effective classroom instruction: Implications of child characteristic by instruction interactions on first graders' word reading achievement. *Journal of Research on Educational Effectiveness*, *4*(3), 173–207. https://doi.org/10.1080/19345747.2010.510179

Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., Underwood, P., & Morrison, F. J. (2009). Individualizing student instruction precisely: Effects of Child × Instruction interactions on first graders' literacy development. *Child Development*, *80*(1), 77–100. https://doi.org/10.1111/j.1467-8624.2008.01247.x

Connor, C. M., Ponitz, C. C., Phillips, B. M., Travis, Q. M., Glasney, S., & Morrison, F. J. (2010). First graders' literacy and self-regulation gains: The effect of individualizing student instruction. *Journal of School Psychology*, *48*(5), 433–455. https://doi.org/10.1016/j.jsp.2010.06.003

Connor, C. M., Radach, R., Vorstius, C., Day, S. L., McLean, L., & Morrison, F. J. (2015). Individual differences in fifth graders' reading and language predict their comprehension monitoring development: An eye-movement study. *The official journal of the Society for the Scientific Study of Reading*, *19*(2), 114–134. https://doi.org/10.1080/10888438.2014.943905

Crevecoeur, Y. C., Coyne, M. D., & McCoach, D. B. (2013). English language learners and English-only learners' response to direct vocabulary instruction. *Reading & Writing Quarterly*, *30*(1), 51–78. https://doi.org/10.1080/10573569.2013.758943

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*(11), 671–684. https://doi.org/10.1037/h0043943

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*, *33*(6), 934–945. https://doi.org/10.1037/0012-1649.33.6.934

Fishman, B. J., Marx, R., Blumenfeld, P., Krajcik, J. S., & Soloway, E. (2004). Creating a framework for research on systemic technology innovations. *Journal of the Learning Sciences*, *13*(1), 43–76. https://doi.org/10.1207/s15327809jls1301_3

Fixsen, D., Blase, K., Metz, A., & Dyke, M. V. (2013). Statewide implementation of evidence-based programs. *Exceptional Children*, *79*(2), 213–230. https://doi.org/10.1177/0014402913079002071

Florit, E., & Cain, K. (2011). The simple view of reading: Is it valid for different types of alphabetic orthographies? *Educational Psychology Review*, *23*(4), 553–576. https://doi.org/10.1007/s10648-011-9175-6

Francis, D. J., Lesaux, N., & August, D. (2006). Language of instruction. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners* (pp. 365–413). Erlbaum.

Fuchs, L. S., Fuchs, D., & Phillips, N. (1994). The relation between teachers' beliefs about the importance of good student work habits, teacher planning, and student achievement. *The Elementary School Journal*, *94*(3), 331–345. https://doi.org/10.1086/461770

Gersten, R., & Baker, S. (2000). What we know about effective instructional practices for English-language learners. *Exceptional Children*, *66*(4), 454–470. https://doi.org/10.1177/001440290006600402

Giambo, D. A., & McKinney, J. D. (2004). The effects of a phonological awareness intervention on the oral English proficiency of Spanish-speaking kindergarten children. *TESOL Quarterly*, *38*(1), 95–117. https://doi.org/10.2307/3588260

Gunn, B., Biglan, A., Smolkowski, K., & Ary, D. (2000). The efficacy of supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school. *The Journal of Special Education*, *34*(2), 90–103. https://doi.org/10.1177/002246690003400204

Haager, D., & Windmueller, M. P. (2001). Early reading intervention for English language learners at-risk for learning disabilities: Student and teacher outcomes in an urban school. *Learning Disability Quarterly*, *24*(4), 235–250. https://doi.org/10.2307/1511113

Hammer, C. S., Jia, G., & Uchikoshi, Y. (2011). Language and literacy development of dual language learners growing up in the United States: A call for research. *Child Development Perspectives*, *5*(1), 4–9. https://doi.org/10.1111/j.1750-8606.2010.00140.x

Hantula, D. A. (2019). Replication and reliability in behavior science and behavior analysis: A call for a conversation. *Perspectives on Behavior Science*, *42*(1), 1–11. https://doi.org/10.1007/s40614-019-00194-2

Hirsch, E. D. (2006). Building knowledge: The case for bringing content into the language arts block and for a knowledge-rich curriculum core for all children. *American Educator*, *30*(1), 8–21.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*(2), 127–160. https://doi.org/10.1007/BF00401799

Kamhi, A. G., & Catts, H. W. (2017). Epilogue: Reading comprehension is not a single ability—Implications for assessment and instruction. *Language, Speech, and Hearing Services in Schools*, *48*(2), 104–107. https://doi.org/10.1044/2017_LSHSS-16-0049

Kamps, D., Abbott, M., Greenwood, C., Arreaga-Mayer, C., Wills, H., Longstaff, J., Culpepper, M., & Walton, C. (2007). Use of evidence-based, small-group reading instruction for English language learners in elementary grades: Secondary-tier intervention. *Learning Disability Quarterly*, *30*(3), 153–168. https://doi.org/10.2307/30035561

Kim, Y.-S. (2017). Why the simple view of reading is not simplistic: Unpacking component skills of reading using a direct and indirect effect model of reading (DIER). *Scientific Studies of Reading*, *21*(4), 310–333. https://doi.org/10.1080/10888438.2017.1291643

Kim, Y.-S. G. (2020). Hierarchical and dynamic relations of language and cognitive skills to reading comprehension: Testing the direct and indirect effects model of reading (DIER). *Journal of Educational Psychology*, *112*(4), 667–684. https://doi.org/10.1037/edu0000407

Kramer, V. R., Schell, L. M., & Rubison, R. M. (1983). Auditory discrimination training in English of Spanish-speaking children. *Reading Improvement*, *20*(3), 162–168.

Lesaux, N. K. (2006). Building consensus: Future directions for research on English language learners at risk for learning difficulties. *Teachers College Record*, *108*(11), 2406–2438. https://doi.org/10.1111/j.1467-9620.2006.00787.x

Lesaux, N., & Geva, E. (2006). Synthesis: Development of literacy in second-language learners. In D. L. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth* (pp. 53–74). Erlbaum.

MacGinitie, W. H., & MacGinitie, R. K. (2006). *Gates-MacGinitie reading tests* (4th ed.). Houghton Mifflin.

Mancilla-Martinez, J., & Lesaux, N. K. (2010). Predictors of reading comprehension for struggling readers: The case of Spanish-speaking language minority learners. *Journal of Educational Psychology*, *102*(3), 701–711. https://doi.org/10.1037/a0019135

Mancilla-Martinez, J., & Lesaux, N. K. (2017). Early indicators of later English reading comprehension outcomes among children from Spanish-speaking homes. *Scientific Studies of Reading*, *21*(5), 428–448. https://doi.org/10.1080/10888438.2017.1320402

Moullin, J. C., Dickson, K. S., Stadnick, N. A., Albers, B., Nilsen, P., Broder-Fingert, S., Mukasa, F., & Aarons, G. A. (2020). Ten recommendations for using implementation frameworks in research and practice. *Implementation Science Communications*, *1*(1), 1–12.

Nakamoto, J., Lindsey, K. A., & Manis, F. R. (2007). A longitudinal analysis of English language learners' word decoding and reading comprehension. *Reading and Writing*, *20*(7), 691–719. https://doi.org/10.1007/s11145-006-9045-7

O'Reilly, T., Sabatini, J., & Deane, P. (2013). *Preliminary reading research assessment framework: Foundation and rationale for assessment and system design (ETS Research Report)*. Educational Testing Service. https://www.ets.org/research/topics/reading_for_understanding/assessments/gisa_samples/

Odom, S. L., Hall, L. J., & Suhrheinrich, J. (2020). Implementation science, behavior analysis, and supporting evidence-based practices for individuals with autism. *European Journal of Behavior Analysis*, *21*(1), 55–73. https://doi.org/10.1080/15021149.2019.1641952

Ouellette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology*, *98*(3), 554–566. https://doi.org/10.1037/0022-0663.98.3.554

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly*, *40*(2), 184–202. https://doi.org/10.1598/RRQ.40.2.3

Pearson, P. D., Palincsar, A. S., Afflerbach, P., Cervetti, G. N., Kendeou, P., Biancarosa, G., Higgs, J., Fitzgerald, M., & Berman, A. I. (2020). Taking stock of the Reading for Understanding initiative. In P. D. Pearson, A. S. Palincsar, G. Biancarosa & A. I. Berman (Eds.), *Reaping the rewards of the Reading for Understanding initiative* (pp. 251–92). National Academy of Education.

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reisma (Eds.), *Precursors of functional literacy* (pp. 189–213). John Benjamins Publishing Company. https://doi.org/10.1075/swll.11.14per

Proctor, C. P., August, D., Carlo, M., & Snow, C. E. (2006). The intriguing role of Spanish vocabulary knowledge in predicting English reading comprehension. *Journal of Educational Psychology*, *98*(1), 159–169. https://doi.org/10.1037/0022-0663.98.1.159

Roehrig, A. D., Duggar, S. W., Moats, L. C., Glover, M., & Mincey, B. (2008). When teachers work to use progress monitoring data to inform literacy instruction: Identifying potential supports and challenges. *Remedial and Special Education*, *29*(6), 364–382. https://doi.org/10.1177/0741932507314021

Scarborough, H. S. (2001). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (pp. 97–110). Guilford Press.

Shadish, W. R., Cook, T. D., & Campbell, J. R. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Company.

Shanahan, T., & Beck, I. L. (2006). Effective literacy teaching for English-language learners. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth* (pp. 415–488). Erlbaum Publishers.

Snow, C. E., & Matthews, T. J. (2016). Reading and language in the early grades. *The Future of Children*, *26*(2), 57–74. https://doi.org/10.1353/foc.2016.0012

Solari, E. J., Terry, N. P., Gaab, N., Hogan, T. P., Nelson, N. J., Pentimonti, J. M., Petscher, Y., & Sayko, S. (2020). Translational science: A roadmap for the science of reading. *Reading Research Quarterly*, *55*(S1), S347–S360. https://doi.org/10.35542/osf.io/8z7e6

Supplee, L. H., & Metz, A. (2015). Opportunities and challenges in evidence-based social policy. *Social Policy Report*, *28*(4), 3–19. https://doi.org/10.1002/j.2379-3988.2015.tb00081.x

Thomas, E., & Sénéchal, M. (2004). Long-term association between articulation quality and phoneme sensitivity: A study from age 3 to age 8. *Applied Psycholinguistics*, *25*(4), 513–541. https://doi.org/10.1017/S0142716404001250

Vaughn, S., Mathes, P. G., Linan-Thompson, S., & Francis, D. J. (2005). Teaching English language learners at risk for reading disabilities to read: Putting research into practice. *Learning Disabilities Research & Practice*, *20*(1), 58–67. https://doi.org/10.1111/j.1540-5826.2005.00121.x

Wanzek, J., Wexler, J., Vaughn, S., & Ciullo, S. (2010). Reading interventions for struggling readers in the upper elementary grades: A synthesis of 20 years of research. *Reading and Writing*, *23*(8), 889–912. https://doi.org/10.1007/s11145-009-9179-5

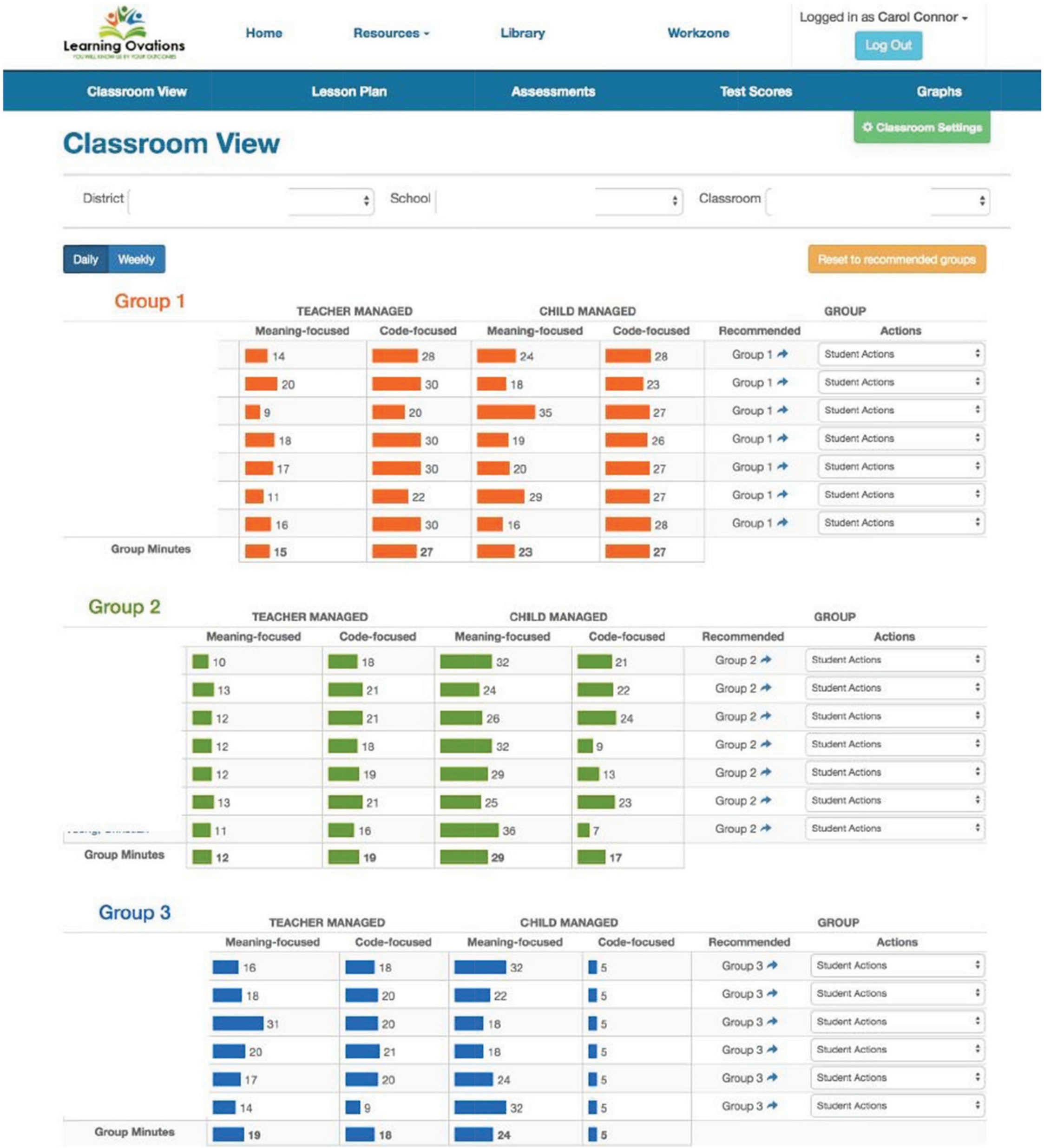
Willingham, D. T. (2006). How knowledge helps: It speeds and strengthens reading comprehension, learning-and thinking. *American Educator*, *30*(1), 30–37.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson-III tests of achievement*. Riverside.

Appendix A

Screenshots of the A2i Technology

Figure A1
Classroom View



Classroom View

Group 1

	TEACHER MANAGED		CHILD MANAGED		GROUP	
	Meaning-focused	Code-focused	Meaning-focused	Code-focused	Recommended	Actions
	14	28	24	28	Group 1 ➦	Student Actions
	20	30	18	23	Group 1 ➦	Student Actions
	9	20	35	27	Group 1 ➦	Student Actions
	18	30	19	26	Group 1 ➦	Student Actions
	17	30	20	27	Group 1 ➦	Student Actions
	11	22	29	27	Group 1 ➦	Student Actions
	16	30	16	28	Group 1 ➦	Student Actions
Group Minutes	15	27	23	27		

Group 2

	TEACHER MANAGED		CHILD MANAGED		GROUP	
	Meaning-focused	Code-focused	Meaning-focused	Code-focused	Recommended	Actions
	10	18	32	21	Group 2 ➦	Student Actions
	13	21	24	22	Group 2 ➦	Student Actions
	12	21	26	24	Group 2 ➦	Student Actions
	12	18	32	9	Group 2 ➦	Student Actions
	12	19	29	13	Group 2 ➦	Student Actions
	13	21	25	23	Group 2 ➦	Student Actions
	11	16	36	7	Group 2 ➦	Student Actions
Group Minutes	12	19	29	17		

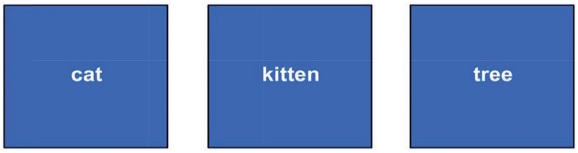
Group 3

	TEACHER MANAGED		CHILD MANAGED		GROUP	
	Meaning-focused	Code-focused	Meaning-focused	Code-focused	Recommended	Actions
	16	18	32	5	Group 3 ➦	Student Actions
	18	20	22	5	Group 3 ➦	Student Actions
	31	20	18	5	Group 3 ➦	Student Actions
	20	21	18	5	Group 3 ➦	Student Actions
	17	20	24	5	Group 3 ➦	Student Actions
	14	9	32	5	Group 3 ➦	Student Actions
Group Minutes	19	18	24	5		

Note. Children's names have been whited out to preserve confidentiality. Each line represents the individual recommendations for one student. See the online article for the color version of this figure.

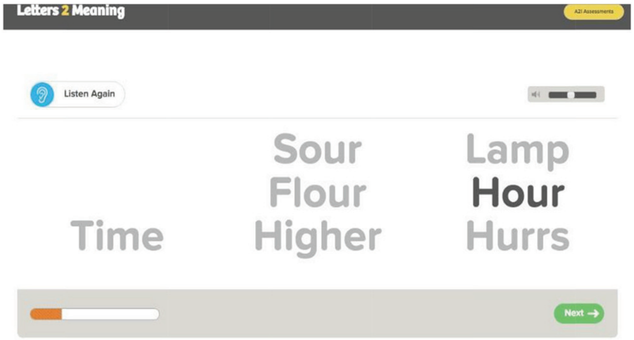
(*Appendices continue*)

Figure A2
Training Item From Word Match Game (WMG)



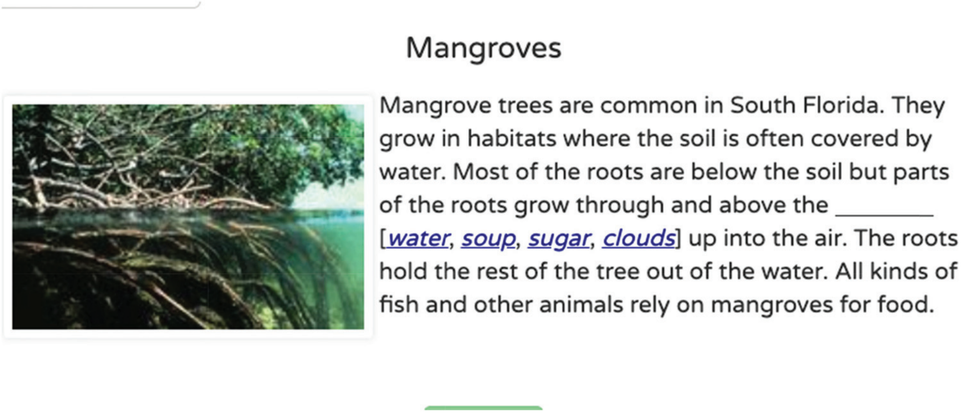
Note. Student hears, "click on the two words that go together." Each word is highlighted as it is said. See the online article for the color version of this figure.

Figure A3
Item From Letters-2-Meaning (L2M)



Note. Student hears "click on the word hour." See the online article for the color version of this figure.

Figure A4
Training Item From Reading2Comprehension (R2C)



Mangroves

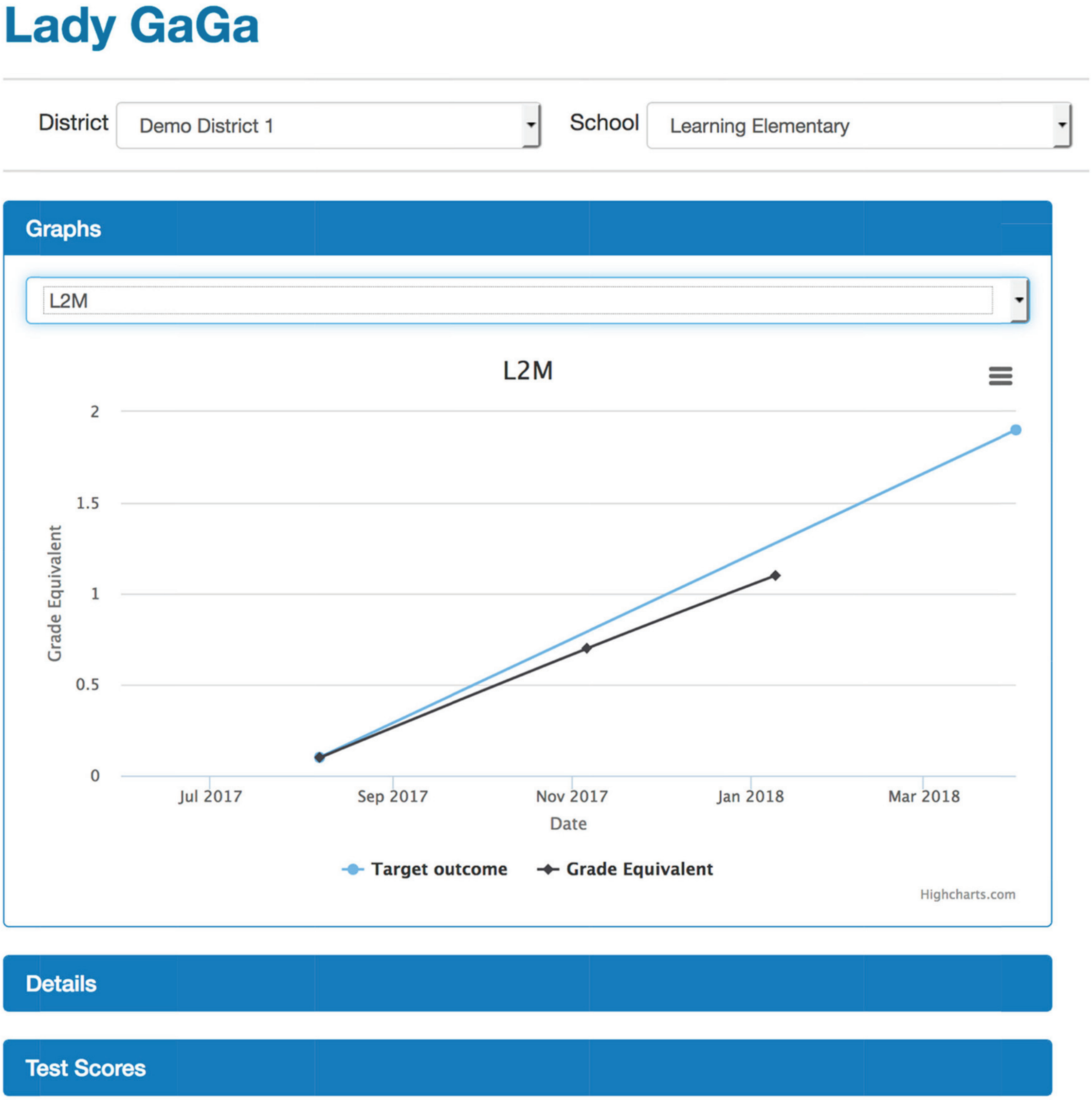
Mangrove trees are common in South Florida. They grow in habitats where the soil is often covered by water. Most of the roots are below the soil but parts of the roots grow through and above the _________ [water, soup, sugar, clouds] up into the air. The roots hold the rest of the tree out of the water. All kinds of fish and other animals rely on mangroves for food.

Note. Students are asked to read passages and choose the best word to fill in the blanks. The instruction and passages are read out aloud for them. See the online article for the color version of this figure.

(*Appendices continue*)

Figure A5

Progress Graph for Individual Student (Not a Real Name)



Lady GaGa

District: Demo District 1

School: Learning Elementary

Graphs

L2M

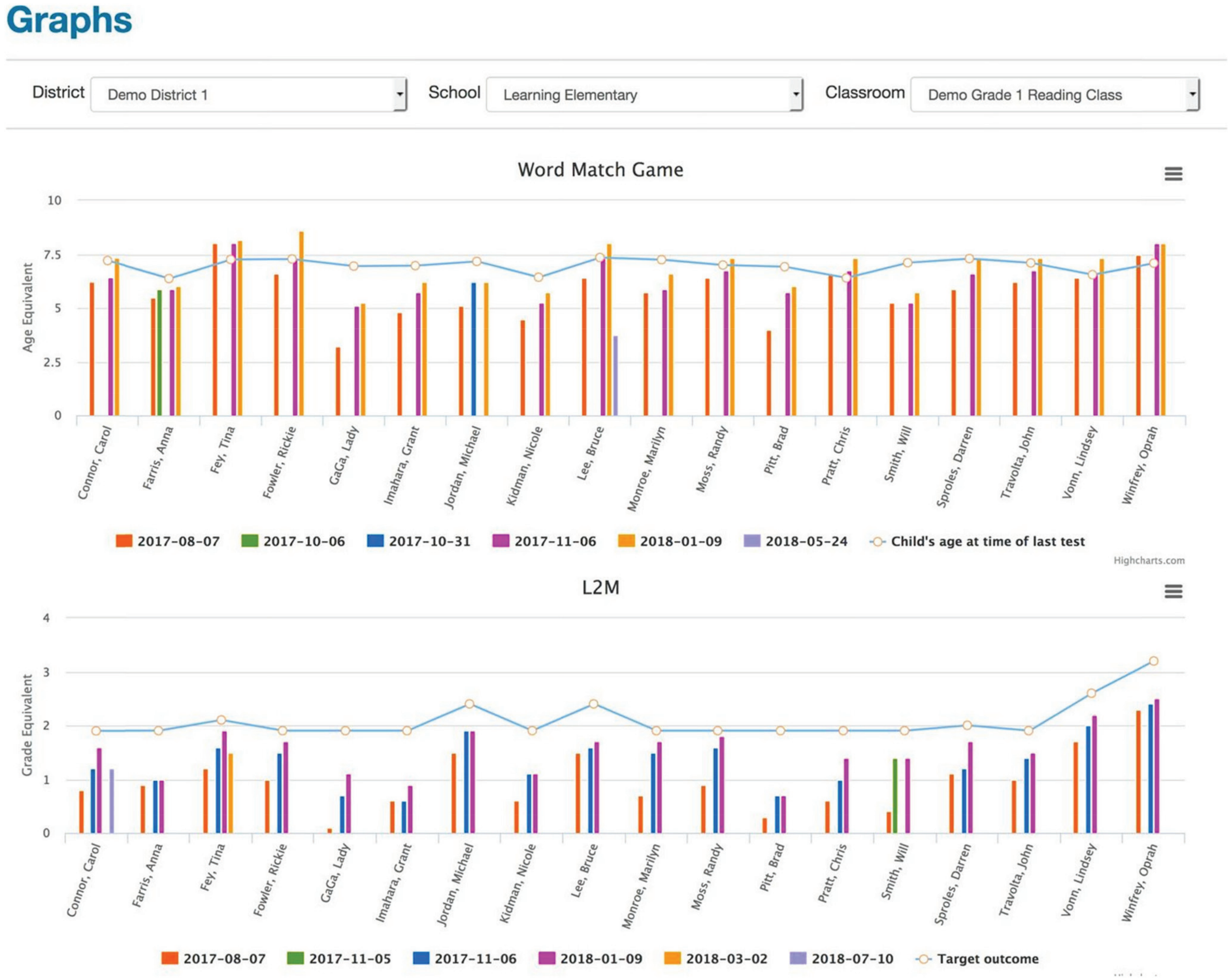
Details

Test Scores

Note. L2M = Letters to Meaning. The blue line represents the target for achievement and the black line shows students' actual progress. See the online article for the color version of this figure.

(*Appendices continue*)

Figure A6
Classroom Graphs



Graphs

District: Demo District 1

School: Learning Elementary

Classroom: Demo Grade 1 Reading Class

Word Match Game

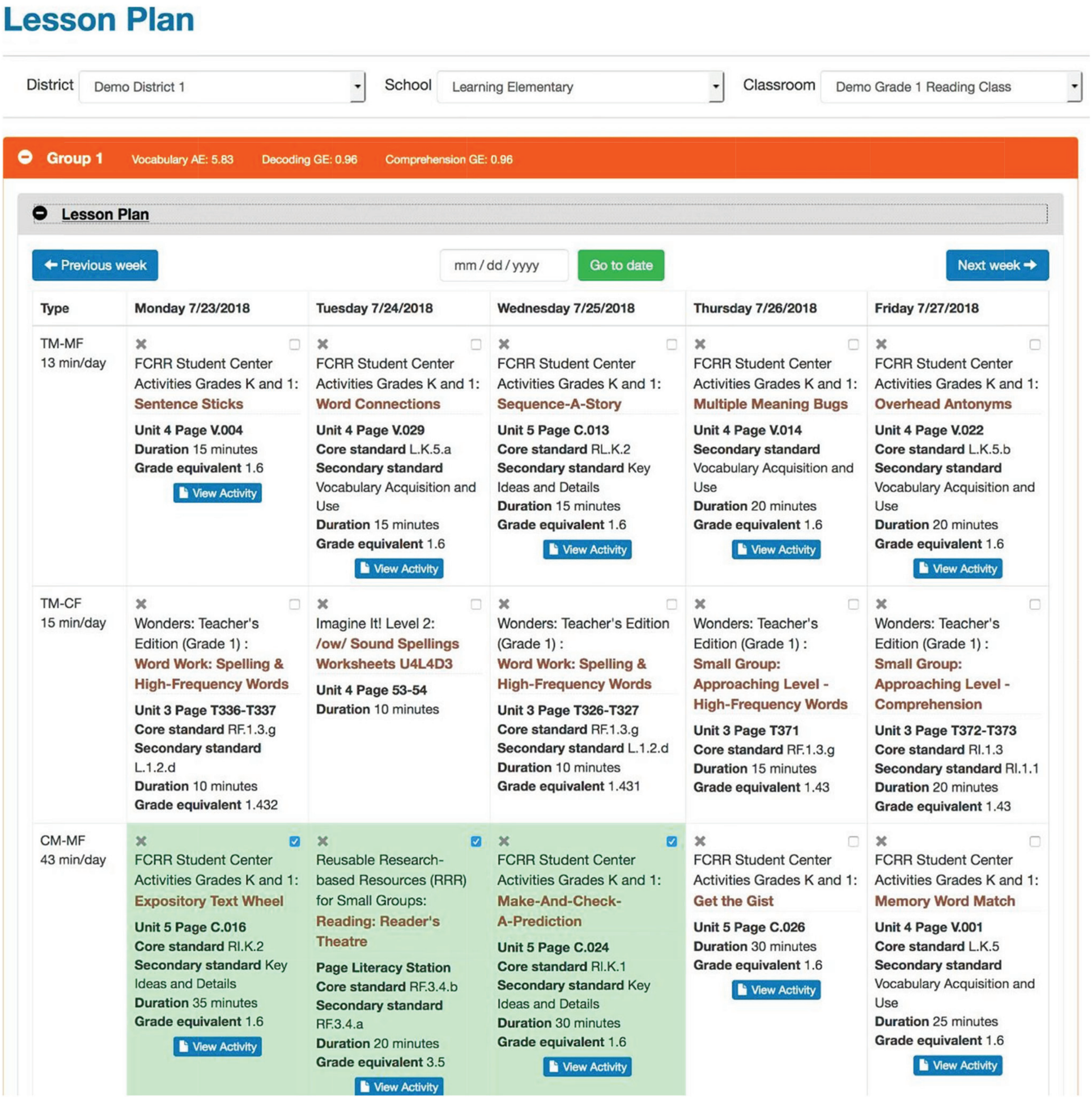
L2M

Note. L2M = Letters to Meaning. Student names are pseudonyms to preserve confidentiality. Each set of bars represents achievement over time for one student. See the online article for the color version of this figure.

(*Appendices continue*)

Figure A7

Lesson Plan Page



Lesson Plan

District: Demo District 1 | School: Learning Elementary | Classroom: Demo Grade 1 Reading Class

Group 1 Vocabulary AE: 5.83 Decoding GE: 0.96 Comprehension GE: 0.96

Lesson Plan

← Previous week | mm / dd / yyyy | Go to date | Next week →

Type	Monday 7/23/2018	Tuesday 7/24/2018	Wednesday 7/25/2018	Thursday 7/26/2018	Friday 7/27/2018
TM-MF 13 min/day	✖ □ FCRR Student Center Activities Grades K and 1: Sentence Sticks **Unit 4 Page V.004** **Duration** 15 minutes **Grade equivalent** 1.6 View Activity	✖ □ FCRR Student Center Activities Grades K and 1: Word Connections **Unit 4 Page V.029** **Core standard** L.K.5.a **Secondary standard** Vocabulary Acquisition and Use **Duration** 15 minutes **Grade equivalent** 1.6 View Activity	✖ □ FCRR Student Center Activities Grades K and 1: Sequence-A-Story **Unit 5 Page C.013** **Core standard** RL.K.2 **Secondary standard** Key Ideas and Details **Duration** 15 minutes **Grade equivalent** 1.6 View Activity	✖ □ FCRR Student Center Activities Grades K and 1: Multiple Meaning Bugs **Unit 4 Page V.014** **Secondary standard** Vocabulary Acquisition and Use **Duration** 20 minutes **Grade equivalent** 1.6 View Activity	✖ □ FCRR Student Center Activities Grades K and 1: Overhead Antonyms **Unit 4 Page V.022** **Core standard** L.K.5.b **Secondary standard** Vocabulary Acquisition and Use **Duration** 20 minutes **Grade equivalent** 1.6 View Activity
TM-CF 15 min/day	✖ □ Wonders: Teacher's Edition (Grade 1) : Word Work: Spelling & High-Frequency Words **Unit 3 Page T336-T337** **Core standard** RF.1.3.g **Secondary standard** L.1.2.d **Duration** 10 minutes **Grade equivalent** 1.432	✖ □ Imagine It! Level 2: /ow/ Sound Spellings Worksheets U4L4D3 **Unit 4 Page 53-54** **Duration** 10 minutes	✖ □ Wonders: Teacher's Edition (Grade 1) : Word Work: Spelling & High-Frequency Words **Unit 3 Page T326-T327** **Core standard** RF.1.3.g **Secondary standard** L.1.2.d **Duration** 10 minutes **Grade equivalent** 1.431	✖ □ Wonders: Teacher's Edition (Grade 1) : Small Group: Approaching Level - High-Frequency Words **Unit 3 Page T371** **Core standard** RF.1.3.g **Duration** 15 minutes **Grade equivalent** 1.43	✖ □ Wonders: Teacher's Edition (Grade 1) : Small Group: Approaching Level - Comprehension **Unit 3 Page T372-T373** **Core standard** RI.1.3 **Secondary standard** RI.1.1 **Duration** 20 minutes **Grade equivalent** 1.43
CM-MF 43 min/day	✖ ☑ FCRR Student Center Activities Grades K and 1: Expository Text Wheel **Unit 5 Page C.016** **Core standard** RI.K.2 **Secondary standard** Key Ideas and Details **Duration** 35 minutes **Grade equivalent** 1.6 View Activity	✖ ☑ Reusable Research-based Resources (RRR) for Small Groups: Reading: Reader's Theatre **Page Literacy Station** **Core standard** RF.3.4.b **Secondary standard** RF.3.4.a **Duration** 20 minutes **Grade equivalent** 3.5 View Activity	✖ ☑ FCRR Student Center Activities Grades K and 1: Make-And-Check-A-Prediction **Unit 5 Page C.024** **Core standard** RI.K.1 **Secondary standard** Key Ideas and Details **Duration** 30 minutes **Grade equivalent** 1.6 View Activity	✖ □ FCRR Student Center Activities Grades K and 1: Get the Gist **Unit 5 Page C.026** **Duration** 30 minutes **Grade equivalent** 1.6 View Activity	✖ □ FCRR Student Center Activities Grades K and 1: Memory Word Match **Unit 4 Page V.001** **Core standard** L.K.5 **Secondary standard** Vocabulary Acquisition and Use **Duration** 25 minutes **Grade equivalent** 1.6 View Activity

Note. See the online article for the color version of this figure.

(*Appendices continue*)

Appendix B

Rubric of Teacher Uptake of Professional Development

Item	Score
Teacher response and participation in communities of practice (COP) meetings (1 = *poor*; 5 = *strong*)	
Teacher attendance in COP (1 = *missed > 2 session*; 5 = *attended all sessions*)	
Teacher response and participation in in-classroom PD (1 = *poor*; 5 = *strong*)	
Teacher attendance in in-class PD (1 = *not willing to schedule*; 3 = *scheduled but ignored Research Partner*; 5 = *scheduled and used feedback*	
Teacher comfort with technology (1 = *not at all comfortable*; 5 = *very comfortable*)	
Teacher feedback on user interface (1 = *not useful*; 5 = *very useful*)	
Teacher willingness to learn how to use A2i (1 = *not willing*; 5 = *very willing*)	
Teacher willingness to meet with Research Partner on a one-to-one basis (1 = *not willing to schedule*, 3 = *scheduled but ignored feedback*, 5 = *scheduled and used feedback*)	
TOTAL	

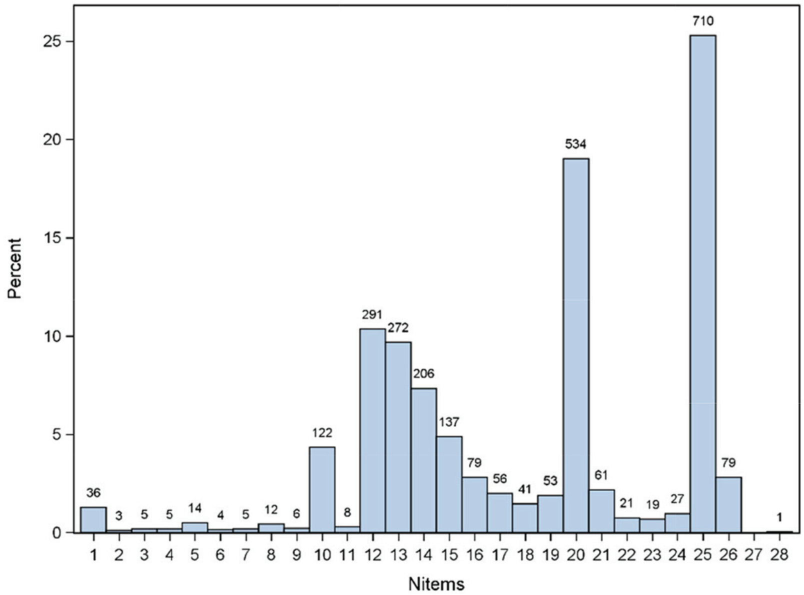
Appendix C

Psychometric Properties for A2i Adaptive Assessments

Scaling Results for the A2i L2M Assessment

A total of 2,807 test administrations were used in the scaling analysis for L2M. Given that the L2M was a computer adaptive assessment, the number of items administered to each student varied. Nearly all test administrations included more than 10 items, and the majority of students responded to 20 or 25 items from the L2M item pool.

Figure C1
Number of Items Administered for L2M Assessment



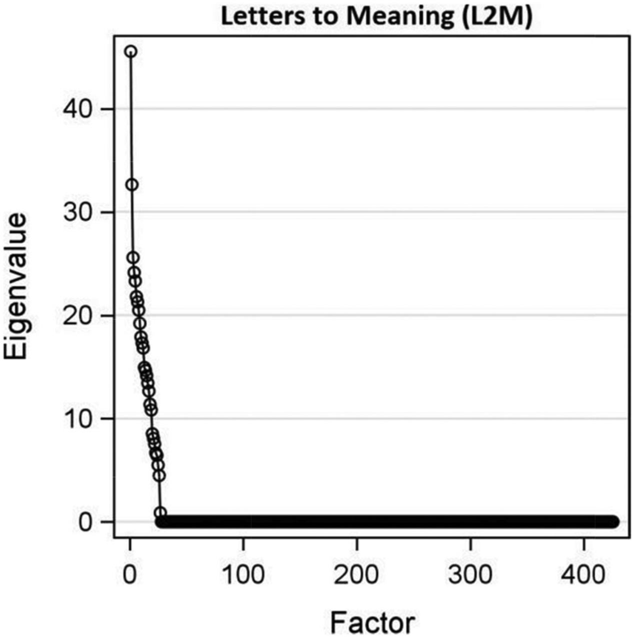
Note. L2M = Letters to Meaning. See the online article for the color version of this figure.

(*Appendices continue*)

Dimensionality

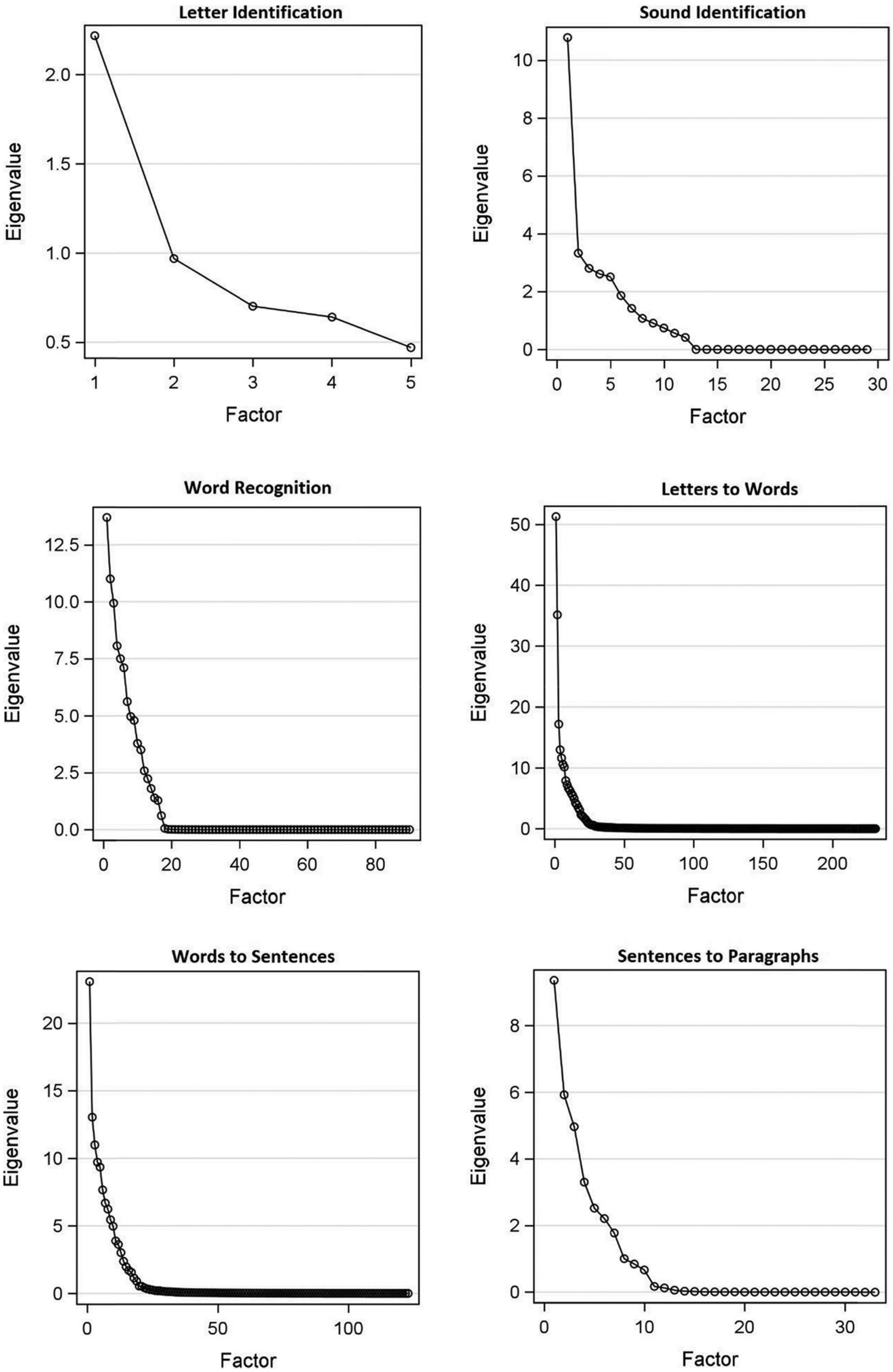
A Scree plot suggests that the L2M assessment is not purely unidimensional. Although the first factor is large, there exists the potential for several subscales. Subsequent analyses were conducted for the overall L2M score, two subscales (i.e., Decoding and Comprehension), and separately for all six subtests within the L2M (that is, Letter Identification (LID), Sound Identification (SID), Word Recognition (WR), Letters to Words (L2W), Words to Sentences (W2S), and Sentences to Paragraphs (S2P).

Figure C2
Dimensionality of the L2M Assessment



(*Appendices continue*)

Figure C3
Dimensionality of L2M Assessment: Specific Subtests



Note. L2M = Letters to Meaning.

(*Appendices continue*)

Item Statistics

Of the 686 items in the L2M item pool, 505 items had more than 30 responses and were included in the Rasch analyses. The average proportion correct across the items was .53 and the median proportion correct was .58 across all items. Item difficulty parameter estimates for the 505 items ranged from −6.5 to +9.3 with a mean difficulty of −.03, a median difficulty of −.21, and a standard deviation of 2.8 points on the Rasch Theta scale. Standard errors for the difficulty estimates ranged from .12 to 1.79 with a mean standard error of .36 and a median standard error of .32 points on the Rasch Theta scale. Full details of item statistics for all L2M items are included in Table A1.

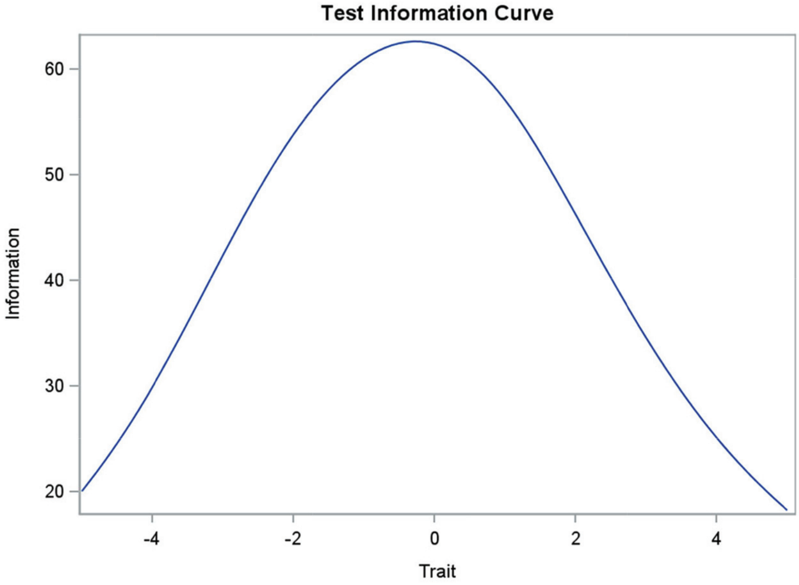
Goodness of Fit

Of the 505 items included in the Rasch scaling, 30 items had more than 200 responses, allowing calculation of an item fit χ^2 statistic. Of these 30 items, only two items had significant goodness of fit statistics ($p < .05$). Both of these items had more than 350 responses, suggesting that the significant χ^2 was a result of a small deviation from the expected values. Inspection of item characteristic curves relative to observed proportion correct confirmed reasonably good fit to the Rasch model despite the significant goodness of fit statistic for these two items.

Test Information

Overall test information for the complete pool of 505 Rasch-scaled L2M items was excellent, with a bell-shaped information function and Total Information greater than 2.0 throughout the range of Rasch theta scores from −5.0 to +5.0, suggesting that computer adaptive administration of L2M will produce reliable individual scores throughout the full range of student abilities.

Figure C4
Test Information for the L2M Assessment



Note. L2M = Letters to Meaning. See the online article for the color version of this figure.

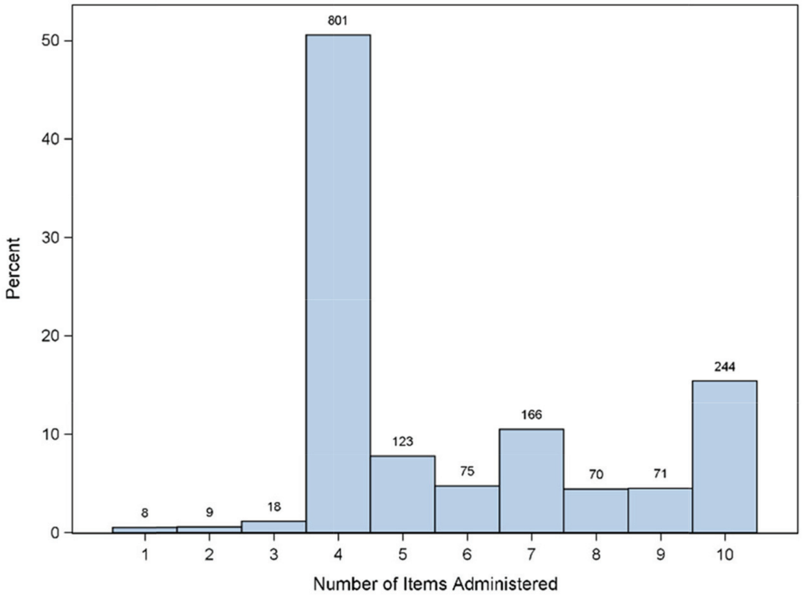
Scaling Results for the A2i R2C Assessment

A total of 1,585 test administrations were used in the scaling analysis for R2C. Given that the R2C was a computer adaptive assessment, the number of items administered to each student varied. Just over half of the test administrations (51%) included four items, 32% included five to nine items, and 15% included all 10 items.

(*Appendices continue*)

Figure C5

Number of Items Administered for R2C Assessment



Note. R2C = Reading to Comprehension. See the online article for the color version of this figure.

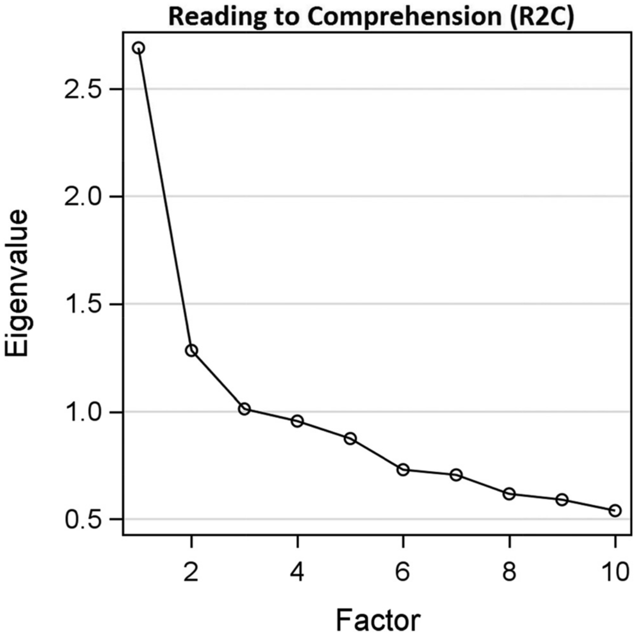
Number of Items Administered for R2C Assessment

Dimensionality

A Scree plot suggests that the R2C assessment is unidimensional. The eigenvalue for the first factor is more than two times larger than the second factor, and the next eight eigenvalues diminish gradually toward zero. This suggests a strong general factor and unidimensionality.

Figure C6

Dimensionality of the R2C Assessment



(*Appendices continue*)

Item Statistics

All 10 items in the R2C item pool had more than 30 responses and were included in the Rasch analyses. The average proportion correct across the items was .37 and the median proportion correct was .32 across all items. Item difficulty parameter estimates for the 10 items ranged from −1.5 to +2.3 with a mean difficulty of +1.28, a median difficulty of +1.53, and a standard deviation of 1.03 points on the Rasch Theta scale. Standard errors for the difficulty estimates ranged from .07 to .16 with a mean standard error of .11 and a median standard error of .10 points on the Rasch Theta scale. Full details of item statistics for all R2C items are included in Table A2.

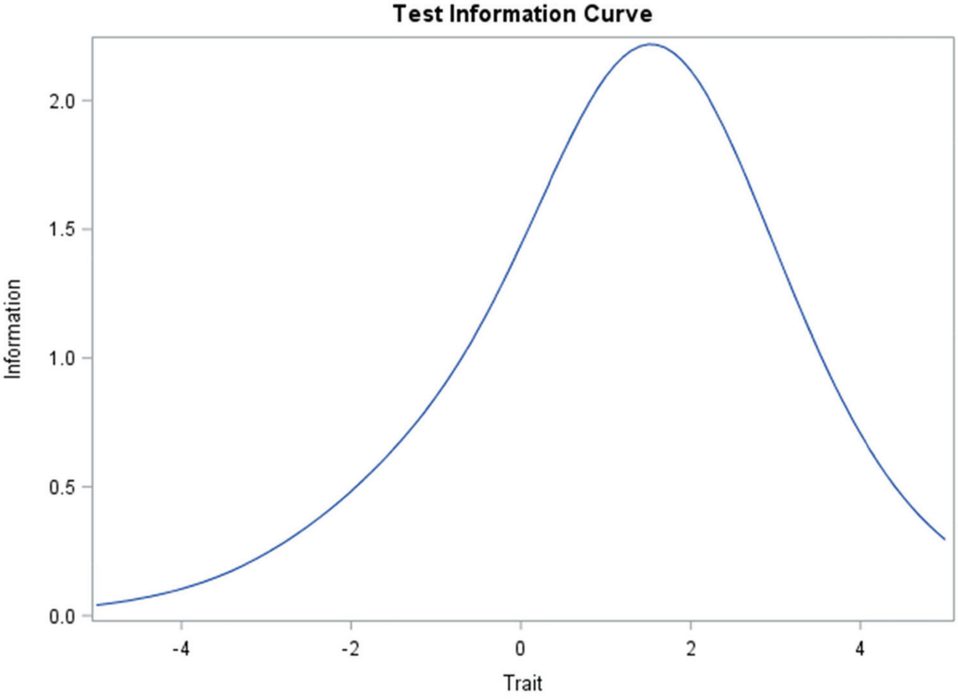
Goodness of Fit

All 10 items included in the Rasch scaling had more than 200 responses, allowing calculation of an item fit χ^2 statistic. None had significant goodness of fit statistics ($p < .05$).

Test Information

Overall test information for the complete pool of 10 Rasch-scaled R2C items was modest, with a bell-shaped information function and Total Information greater than 2.0 for Rasch theta scores in the range +1.0 to +3.0, suggesting that computer adaptive administration of R2C will produce reliable individual scores only in the upper range of student abilities and that reliability of R2C scores at the lower end would be improved if additional items were added to the R2C item pool.

Figure C7
Test Information for the R2C Assessment



Note. R2C = Reading to Comprehension. See the online article for the color version of this figure.

Overall, second and third graders achieved means of 1.32 and 1.47, respectively, with an ICC for student of .17 and for teachers .19. These are not out of line with students' scores on the GM.

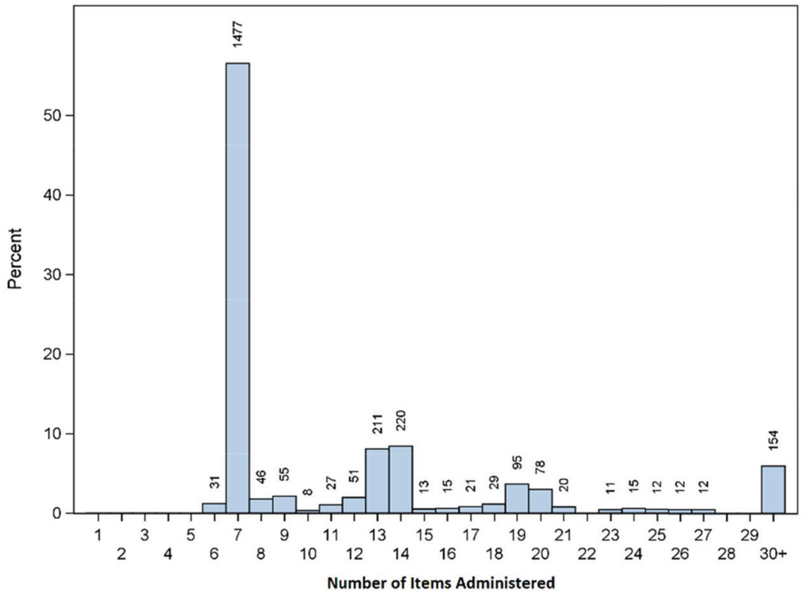
Scaling Results for the A2i WMG Assessment

A total of 2,613 test administrations were used in the scaling analysis for WMG. Given that the WMG was a computer adaptive assessment, the number of items administered to each student varied. Just over half of the test administrations (57%) included seven items, 36% included eight to 29 items, and 6% included 30 or more items.

(*Appendices continue*)

Figure C8

Number of Items Administered for the WMG Assessment



Note. WMG = Word Match Game. See the online article for the color version of this figure.

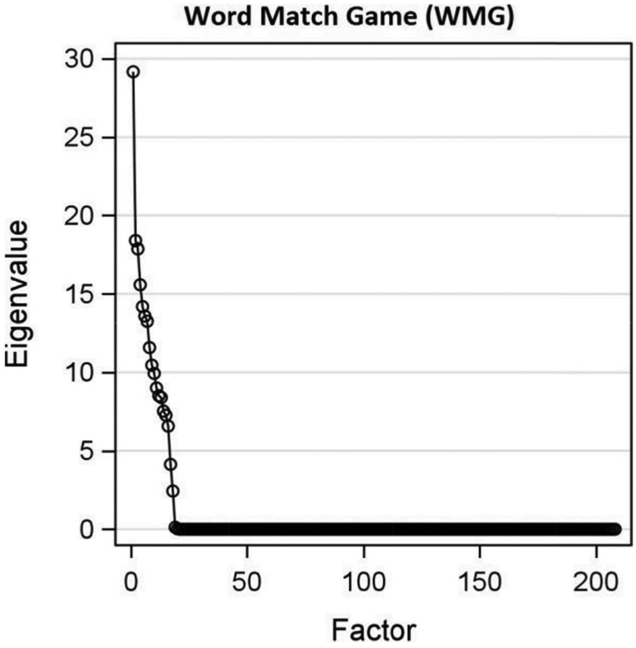
Number of Items Administered for WMG Assessment

Dimensionality

A Scree plot suggests that the WMG assessment may be unidimensional. The eigenvalue for the first factor is approximately 1.5 times larger than the second factor, and the next 18 eigenvalues diminish gradually toward zero. Given the large item pool of 209 items and the small number of responses for some items, this suggests a strong general factor and possible unidimensionality.

(*Appendices continue*)

Figure C9
Dimensionality of the WMG Assessment



Item Statistics

All 209 items in the WMG item pool had more than 30 responses and were included in the Rasch analyses. The average proportion correct across the items was .61 and the median proportion correct was .63 across all items. Item difficulty parameter estimates for the 209 items ranged from −3.3 to +3.5 with a mean difficulty of −.38, a median difficulty of −.39, and a standard deviation of 1.03 points on the Rasch Theta scale. Standard errors for the difficulty estimates ranged from .07 to .56 with a mean standard error of .24 and a median standard error of .24 points on the Rasch Theta scale. Full details of item statistics for all WMG items are included in Table A3.

Goodness of Fit

Of the 209 items included in the Rasch scaling, 53 items had more than 200 responses, allowing calculation of an item fit χ^2 statistic. Of these 53 items, none had significant goodness of fit statistics ($p < .05$).

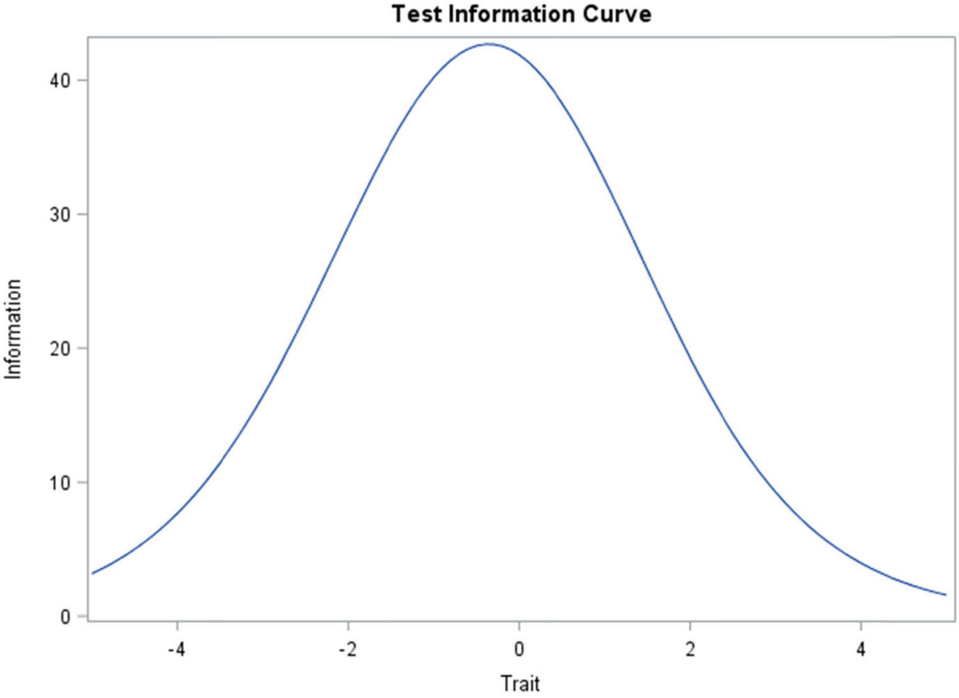
(*Appendices continue*)

Test Information

Overall test information for the complete pool of 209 Rasch-scaled WMG items was excellent, with a bell-shaped information function and Total Information greater than 2.0 throughout the range of Rasch theta scores from −5.0 to +5.0, suggesting that computer adaptive administration of WMG will produce reliable individual scores throughout the full range of student abilities.

Figure C10

Test Information for the WMG Assessment



Note. WMG = Word Match Game. See the online article for the color version of this figure.

Received February 26, 2020
Revision received March 26, 2022
Accepted April 11, 2022 ■