Diagnostic Online Reading Assessment &

DORA – Spanish/EDELL (Evaluación Diagnostica Español de Lectura en Línea):

Technical Document

Published by Let’s Go Learn, Inc.
# Table of Contents

Introduction.......................................................................................................................... 3  
*Diagnostic Online Reading Assessment* Specifications.......................................................... 4  
  High Frequency Words Sub-Test .......................................................................................... 5  
  Phonemic Awareness Sub-Test ............................................................................................ 5  
  Phonics Sub-Test ................................................................................................................ 6  
  Word Recognition Sub-Test ............................................................................................... 7  
  Vocabulary Sub-Test .......................................................................................................... 7  
  Spelling Sub-Test ............................................................................................................... 8  
  Silent Reading Comprehension Sub-Test ........................................................................... 9  
  Fluency Sub-Test ............................................................................................................. 10  
*DORA* – *Spanish/EDELL (Evaluación Diagnostica Español de Lectura en Línea)* Specifications .................................................. 11  
Analyzing Sub-tests: Reports and Instructional Connections ................................................. 13  
*DORA* Content Validity ..................................................................................................... 15  
*DORA* Construct Validity ................................................................................................ 17  
  Tahoe-Truckee Unified Schools (2002) Correlation Study .................................................. 17  
  Item Analysis Major Revision (2003) ............................................................................... 18  
  Recalculation of April SP and SR results & New Tests (2003) ..................................... 19  
  New York City Schools (2009) Correlation Study ............................................................. 20  
  Ault-Highland Schools (2010) Hierarchical Linear Modeling ........................................... 20  
  Jersey City Schools (2013) Correlation and Regression Analysis ..................................... 21  
*DORA* Internal-Consistency Reliability Study .................................................................. 22  
  Introduction ...................................................................................................................... 22  
  Methods ............................................................................................................................ 22  
  Analysis ............................................................................................................................ 23  
  Conclusions ...................................................................................................................... 23  
*DORA Spanish* Internal-Consistency Reliability Study ......................................................... 25  
  Introduction ...................................................................................................................... 25  
  Methods ............................................................................................................................ 25  
  Analysis ............................................................................................................................ 26  
  Conclusions ...................................................................................................................... 26  
*DORA* Test-Retest Reliability Study (2003) ...................................................................... 28  
References ......................................................................................................................... 29  
Appendix A: Silent Reading Leveled Passages Design Specifications ..................................... 31  
  Fact v. Fiction ................................................................................................................... 31  
  Topics ............................................................................................................................... 32  
  Stylistic Design ................................................................................................................ 33  
  Comprehension Questions ............................................................................................... 33  
Appendix B: Grade Level Justifications and Parallels for *DORA* Silent Reading Leveled Passages .................................................. 36  
  Passage Length and Paraphrasing ..................................................................................... 37  
  Sentence Length and Complexity ..................................................................................... 37  
  Vocabulary ....................................................................................................................... 39  
Appendix C: The Relationship between Online Formative Assessment Scores and State Test Scores Using Multilevel Growth Modeling ............................................................... 41  
Appendix D: The predictive values of DORA scores in New Jersey Statewide Assessment (NJ ASK) .......................................................... 51
Introduction

In response to a growing need to increase the literacy achievement of their students, most schools today have adopted programs that utilize a traditional, teacher-centered, whole-class approach, while others have opted to work with small groups and still others have employed one-on-one tutoring for select students. Prior studies have shown that a variation on the latter approach -- individualized instruction based on diagnostic assessment information -- is an ideal way to prevent reading failure and accelerate the literacy development of low-achieving students (Juel, 1996; Shanahan & Barr, 1995; Wasik & Slavin, 1998; McCallum, et al., 2000). Let’s Go Learn’s Diagnostic Online Reading Assessment and LGL Edge instructional programs were created to provide teachers and parents with individualized reading assessment and instruction. The theoretical, pedagogical, and curricular dimensions of the program were based on best practices developed by reading specialists who worked directly with struggling readers in the Cal Reads program at UC Berkeley and who studied under Dr. Richard McCallum, the director of the Advanced Reading and Language Leadership Program at the Graduate School of Education at UC Berkeley.

Although much research has been conducted on the importance of individualized and comprehensive reading instruction, current models of individualized instruction are typically limited to serving only a fraction of the students in need, while the pool of trained professionals necessary to implement these programs is small. Let’s Go Learn has developed a solution to provide individualized literacy assessment and instruction to all students and a way for teachers to manage each student’s reading progress and apply
individualized computer-based activities to authentic classroom practices. Let’s Go Learn’s assessment, *DORA (Diagnostic Online Reading Assessment)*, and individualized online instructional program, *LGL Edge*, were modeled after Cal Reads, a successful program developed by Dr. McCallum at UC Berkeley. In his tutoring program, Dr. McCallum had shown that consistent, individualized reading assessment and instruction could raise literacy skills of struggling readers by approximately two years within a school year as compared to a control group of similar students (McCallum et al., 2000). Furthermore, a study examining the effects of the proposed reading program (Learning Today, 2003) on the reading achievement levels of Title 1/Limited English Proficient (LEP) students in Belle Glade, Florida showed significant grade-level gains in their sight words, phonics, vocabulary, and reading comprehension skills after 12 weeks in the program when compared to the grade-level gains of students in a control group.

*Diagnostic Online Reading Assessment Specifications*

Because a single standardized assessment gives only a small picture of a student’s reading ability and provides little formative information to guide instructional practices, *DORA* was created with the idea of obtaining information from multiple sources and vantage points to more accurately paint a picture of an individual’s reading strategies across multiple measures which follow a constructivist perspective of the reading process (Flores et al., 1991). The most effective way to characterize a child’s reading ability is to assess his or her reading skills across a set of criterion-referenced categories that are important to the reading process. Modeled after the battery of reading measures used by Cal Reads and reading specialists as outlined in various reading methods books (Gillet & Temple, 1994;
Ruddell, 1999), the eight reading skills measured by Let’s Go Learn are: 1) High Frequency Words, 2) Phonemic Awareness, 3) Phonics, 4) Word Recognition, 5) Vocabulary, 6) Spelling, 7) Silent Reading Comprehension, and 8) Fluency.

High Frequency Words Sub-Test

This sub-test assesses children’s ability to automatically recognize words that have been identified as frequently occurring in books, newspapers, and other texts. This sub-test uses words from Edward B. Fry’s 300 sight words as test items which have been broken down into three general levels of difficulty (Fry, Kress, & Fountoukidis, 2004). A child’s response time in identifying these sight words is recorded and factored into the scoring of the child’s performance on the assessment. High-Frequency Word subtest consists of 72 criterion-referenced words, with 24 words per grade from 1st to 3rd grade.

Phonemic Awareness Sub-Test

According to Ruddell (1998), by the time children are between three and four years old, they have learned most of the approximately 40 phonemes (discrete sounds in words) which comprise the English language. The ability to hear and manipulate these discrete sounds in spoken words is referred to as “phonemic awareness.” Children demonstrate their phonemic awareness by segmenting words into individual sounds (i.e., /fish/ into /f/-/i/-/sh/), deleting sounds in words, blending sounds, adding sounds, or substituting sounds within a word to make a new word. Children who have good phonemic awareness can often recognize/decode words and spell/write better than others. Some researchers have indicated that phonemic awareness is one of the best predictors of reading success (Stanovich, 1993-1994). Others further argue that phonemic awareness
is both the prerequisite and consequence of learning to read (Yopp, 1992). As such, it is especially important to determine children’s level of phonemic awareness in the primary grades to ensure that they get any necessary intervention as early readers, lest they struggle with reading as young adults.

On the phonemic awareness sub-test, children are presented with a number of audio and picture-only items and asked to manipulate the sounds in these items to produce a new word. Specific phonemic awareness categories tested include: 1) addition, 2) deletion, 3) substitution, 4) identification, 5) categorization, 6) blending, 7) segmenting, 8) isolation, and 9) rhyming.

Phonics Sub-Test

In addition to having an awareness of the discrete sounds in words, it is important that children have a mastery of how sounds and words are represented in English. This is important because children need to be able to effortlessly decode and recognize familiar and unfamiliar words to help facilitate the process of negotiating the meaning behind the text (Adams, 1990; Snow, Burns, & Griffin, 1998). The phonics sub-test assesses a child’s ability to recognize basic English phonetic principles of high utility (Pressley & Woloshyn, 1995). These phonetic principles include: 1) beginning sounds, 2) short vowel sounds, 3) blends, 4) the silent E rule, 5) consonant digraphs, 6) vowel digraphs, 7) r-controlled vowels, 8) diphthongs, and 9) syllabification. The Phonics subtest consists of 80 criterion-referenced words, with 20 words per grade, from 1st to 4th grade.
Word Recognition Sub-Test

As in many informal reading inventories such as the Qualitative Reading Inventory (Leslie & Caldwell, 1994), the Basic Reading Inventory (Johns, 2001) and the Diagnostic Assessment of Reading (Roswell & Chall, 1992), DORA’s Word Recognition sub-test assesses a learner’s ability to recognize leveled lists of words. In this sub-test, children are presented with a number of increasingly difficult words until they reach a level at which they “frustrate” or stop recognizing the words presented to them. The final outcome of the assessment gives teachers an idea of the grade-level ability of a child to recognize words out of context. This assessment is important in identifying how well an individual can use what he or she knows about text to recognize words outside the context of a sentence and of increasing difficulty. The Word Recognition sub-test consists of 120 criterion-referenced words, with 10 words per grade from 1st to 12th grade.

Vocabulary Sub-Test

A learner’s knowledge of words and what they mean is an important part of the reading process, as knowledge of word meanings affects the extent to which the learner comprehends what he or she reads (National Reading Panel, 2000). The vocabulary sub-test assesses a child’s understanding of words. The words from this sub-test were selected by teachers and reading specialists to reflect the types of words children learn in various disciplines at different grade levels and in various stages of their lives. Similar to the Peabody Picture Vocabulary Test (Dunn, 1959), in the vocabulary sub-test children are asked to select the picture which correctly corresponds to a word they hear. The program continues to present children with increasingly difficult words until they make a certain number of errors. This sub-test provides information about a child’s level of oral
vocabulary. The Oral Vocabulary subtest consists of 60 criterion-referenced words, with 5 words per grade, ranging from 1st to 12th grade.

**Spelling Sub-Test**

The process of spelling involves a number of cognitive processes. While each person uses different strategies for spelling words, these strategies usually have in common a familiarity with a particular word (i.e., familiarity with its meaning and visual exposure to the word), letter-sound matching, and confirmation of how the word “looks” (Bear et al., 2000; Ruddell, 1999; Gillet & Temple, 1994). Because spelling is also a generative process (as opposed to a decoding and meaning-making process in reading), it is natural for young readers’ spelling abilities to lag a few months behind their reading abilities. *DORA’s* Spelling sub-test tries to capture the nuances of the different processes children use to spell words by employing target words with increasing difficulty in different domains. In the process of creating the items for the *DORA* Spelling sub-test, reading specialists created a list of recommended target spelling words by examining words commonly encountered in or taught at particular grade levels. The difficulty of the recommended words changes in these general domains in the following ways: 1) number of syllables in a word, 2) regular phonetic patterns within the words, 3) irregular phonetic patterns within the words, 4) vocabulary level, and 5) a child’s expected familiarity with a word based on his or her grade level. In the first through third grade spelling lists, for instance, while most of the words are phonetically regular, the number of syllables increases with each grade level and the phonetic patterns within each word become increasingly difficult. The way the difficulty increases per grade level varies. Studies of how students perform on the spelling test have been used to eliminate words that are too
easy or too difficult for a particular grade level, resulting in a test that possesses the
ability to better distinguish students who spell well on one list of words as opposed to
another.

This sub-test is important in providing insight about a learner’s orthography skills--
that is, a learner’s ability to take what he or she knows about letters and sounds and
represent it in words and ability to represent words with irregular spelling patterns. This
sub-test measures a child’s conventional spelling ability as well as provides teachers with
information about the errors their students make. The program invites children to correctly
spell a series of words that becomes increasingly difficult. The program stops administering
words when a child consistently spells words incorrectly. Items from this sub-test were
chosen by reading specialists and classroom teachers to approximate the kinds of words
children of a particular age would see in their classroom instruction. The Spelling subtest
consists of 60 criterion-referenced words, with 5 words per grade, ranging from 1st to 12th
grade.

Silent Reading Comprehension Sub-Test

The silent reading comprehension sub-test forms the crux of DORA, which attempts
to provide a window into the semantic domain of a learner’s reading abilities. The content
of each silent reading passage is expository and written to reflect the subject areas that
students of a particular grade level would encounter. In a variation on protocols for some
informal reading inventories (Gillet & Temple, 1994; Leslie & Caldwell, 1994), children
silently read passages of increasing difficulty and answer questions about each passage
immediately after they read it. The questions for each passage are broken up into three
factual questions, two inferential question, and one contextual vocabulary question. The
program stops administering passages and questions once a student misses a certain number of questions on a passage. It provides teachers with information about a child’s comprehension level. For complete technical specifications on the leveled passages, please see Appendix A: Silent Reading Design Specifications and Appendix B: Grade Level Justifications and Parallels. The Reading Comprehension subtest consists of 12 Flesch-Kincaid leveled passages with 6 questions per passage. There is one passage per grade, with three sets of comprehension passages used. Students cycle through A, B, and C passages on subsequent assessments.

*Fluency Sub-Test*

Fluency is included as a teacher-administered measure. In this sub-test, children read aloud to teachers short leveled passages with increasing syntactic complexity. Teachers time children’s reading of these passages and record their errors and prosody (voice inflection, articulation, and versification) according to a pre-established rubric adapted from the National Assessment of Educational Progress (NAEP) Oral Reading Fluency Scale (1995).
DORA – Spanish/EDELL (Evaluación Diagnostica Español de Lectura en Línea) Specifications

Concern about the lack of access to available tools in Spanish from many educators of English Language Learners prompted the development of DORA Spanish/EDELL. As the number of language minority students in school rises, so does the number of language minority students who fall further and further behind their English-speaking peers. To start addressing the needs of Spanish-speaking English Language Learners in their schools, educators have recognized the need to identify both L1 and L2 competencies of Spanish-speaking children to provide the best instructional placement for these children.

According to the National Literacy Panel on Language-Minority Children and Youth (2006), “Language minority students are not blank slates. They enter classroom rooms with varying degrees of oral proficiency and literacy in their first language. There is clear evidence that tapping into first-language literacy can confer advantage.” Tapping into the first-language literacy proficiencies of Spanish-speaking English language learners is the goal of DORA Spanish. It utilizes the same underlying principles as English DORA; that is, reading in Spanish involves similar cognitive strategies and attention to similar sets of domains (albeit with different parameters) when reading and negotiating the meaning of text. As such, to characterize a child’s Spanish reading profile, DORA Spanish examines a child’s reading abilities across similar domains in Spanish: 1) High Frequency Words, 2) Phonemic Awareness, 3) Phonics, 4) Word Recognition, 5) Vocabulary, 6) Spelling, and 7) Silent Reading Comprehension. DORA Spanish takes results from these seven subskills and reports a quantitative and qualitative account of the child’s Spanish reading abilities, just as it does in English.
**DORA** Spanish/EDELL is a powerful tool that should be used to help educators make better instructional decisions for their Spanish-speaking students. **DORA** Spanish will be especially useful in tapping into and capitalizing upon the L1 knowledge of students so that teachers can help accelerate the development of L2 while children maintain fluency in their first language. As the National Literacy Panel on Language-Minority Children and Youth reports, “Research indicates that instructional programs work when they provide opportunities for students to develop proficiency in their first language.” **DORA** Spanish/EDELL is a powerful tool that will help children become truly bi-literate.
Analyzing Sub-tests: Reports and Instructional Connections

After a child completes an assessment, teachers can immediately retrieve a report which details his or her reading profile in a quantitative and qualitative fashion and provides important instructional recommendations specific to that child’s reading profile.

There are multiple components of a teacher’s report on a child’s reading assessment. The first section includes a quantitative summary of the child’s performance on each of the sub-tests. In this section of the report, areas of low performance are flagged to make teachers aware that that particular reading skill is of high priority to that child. The next section of the report includes a qualitative summary of the child’s reading profile which is broken up into three sections: 1) Graphophonic Strategies, 2) Semantic and Syntactic Strategies, and 3) Overall Summary. According to Adams (1998), learners attend to letter/sound (graphophonic), meaning-based (semantic), and language-based (syntactic) cues as they read, sometimes relying more heavily on one cue than on another as they draw from their background knowledge. The three domains (graphophonic, semantic, and syntactic) are used to analyze each learner’s reading profile and to extrapolate, via the results of DORA’s sub-tests, each child’s strengths and weaknesses and the strategies he or she uses when encountering text. The graphophonic section characterizes a child’s attention to letter and sound cues in words. This includes his or her ability to recognize words by sight, application of phonics principles, and ability to distinguish sounds within words. The semantic and syntactic section analyzes the child’s ability to attend to meaning and language-based cues. The overall summary explains the child’s strengths and weaknesses in utilizing all three strategies--graphophonic, semantic, and syntactic--to decode and make sense of what he or she reads. Each section uses evidence from the sub-tests listed above to
qualify the profile made of the student’s reading abilities, including scores and, where appropriate, a description of the child’s errors relative to the target item.

The third section of the report is a summary of different kinds of instructional strategies in the three domains (graphophonic, semantic, and syntactic) which would best accelerate the child’s reading development. The instructional strategies draw from the child’s successes and known abilities to help build the areas of weakness identified by DORA. The last section of the report provides other detailed instructional strategies and tips in teaching reading that would be helpful to all students in the classroom.

To facilitate the use of reading assessment information in meaningful and formative ways, Let’s Go Learn’s DORA provides teachers with an online management system. This online management system allows teachers to sort assessment information in order to group children with similar profiles, monitor individual and classroom progress, and quickly view individual assessment summaries. The online management system assists teachers in their various instructional decisions, including grouping and selecting appropriate reading activities and materials.

All of the sub-tests, except for the fluency sub-test, are conducted online with little teacher mediation. With DORA, a teacher has the capability of assessing her whole classroom in one trip to the school’s computer laboratory. Unlike other assessments, which often require multiple teachers, reading specialists, or elaborate management setups for a period of weeks to assess individual students, DORA provides individualized, diagnostic, and comprehensive reading assessments coupled with a powerful data management and reporting system that helps teachers make appropriate instructional decisions for their classrooms and provides progress information on every child.
**DORA Content Validity**

The validity of an assessment instrument refers to its ability to support valid assessment inferences. That is, do test results support a valid conclusion about a student's level of knowledge or skill? Building a valid test begins with accurate definitions of the content (i.e., the knowledge domains and skills) to be assessed. If the assessment activities in a test (i.e., the test items) tap into the constructs that the test is designed to assess, then the test has content validity. Although additional factors affect overall test validity, construct validity is the basic logical bedrock of any test.

The content validity of Let's Go Learn's reading assessments derives from the most current research-based and classroom-proven models of reading sub-skill acquisition and diagnostic reading assessment. These models are based on the work of Richard McCallum, Ph.D., a recognized expert in using diagnostic assessment to tailor instructional interventions precisely to students' individual profiles of reading sub-skill performance. Dr. McCallum is a co-founder of Let's Go Learn and is the company's Chief Educational Architect.

Richard McCallum has many years' experience training graduate students as reading specialists at Saint Mary's College and at UC Berkeley. During this period, he and his teams of classroom intervention specialists have also trained hundreds of elementary, middle and high school teachers to diagnose sub-skill deficiencies and to turn their at-risk students into strategic readers. He has conducted lectures and training seminars on reading instruction and assessment for district administrators and principals. Dr. McCallum combines superb scholarly and academic credentials with many years of...
practical experience in implementing extremely effective classroom instructional intervention. He is uniquely suited to guide and integrate Let's Go Learn's pioneering efforts in creating and expanding an online educational environment based on valid assessment and effective instruction.
In its first comparison to one-on-one paper-and-pencil assessments performed by CAL Reads reading specialists, Let’s Go Learn achieved high correlations with statistical significance beyond the a= .01 level, indicating another measure of assessment validity (n=20). This study was conducted in the Tahoe/Truckee Unified School District in California in February 2002. Students were tested both by CAL Reads reading specialists and online using Let’s Go Learn within a three-week time period. The Word Analysis subtests were not compared because of incompatible methods with which CAL Reads and Let’s Go Learn reported their final results. See our paper, SBIR U.S. Department of Education Pilot Results for more information surrounding this pilot.

CAL Reads reading specialists used the following assessments in their one-on-one assessments with the students in the study:

- Sight-word familiarity: Fry’s high-frequency word list
- Word recognition: Diagnostic Assessments of Reading (DAR) published by Riverside Publishing
- Word meaning: Diagnostic Assessments of Reading (DAR) published by Riverside Publishing
- Spelling: Diagnostic Assessments of Reading (DAR) published by Riverside Publishing
- Silent Reading: Qualitative Reading Inventory (QRI) published by Pearson, Allyn & Bacon.
  - Sight-word familiarity $r= .89$
  - Word recognition $r= .81$
  - Word meaning $r= .60^*$
• Spelling $r = .78$
• Silent reading $r = .89$

Lower correlation results in Word Meanings were traced to two confusing items in this initial study. Shortly after this initial study, the Word Meaning component of *DORA* was modified to improve its correlation.

*Item Analysis Major Revision (2003)*

Item analysis was performed across all LGL assessment systems. A pool of 1,000 students was used. Items across all six sub-tests with more than 75% of students answering correctly or fewer than 25% answering correctly were flagged. These extreme values represent errors that are outside the range for which the items were designed. Subsequent subtest modifications and item revisions were then made.

Example 1: "and" received an unusually high error rate. The word "an" was a distracter that was selected with a high percentage. Conclusion: The audios of "and" and "an" are too similar. Students might not be hearing the /d/ sound and thus may think the target word is "an."

Example 2: Target vocabulary word: "Caravan." Too often students chose a picture of one car and one van. Overall the error rate was too high for this particular word. Conclusion: Many students define "caravan" as the Dodge Caravan vehicle and not as a line of camels walking through the desert.

Following the item revisions in early 2003, subtest correlation studies were again conducted. Students (grade range: 2-6) were assessed within one week’s time on DORA and the following paper-and-pencil tests.

- DORA HFW subtest and the Slosson Oral Reading Test
- DORA WR subtest and the Woodcock Word Identification Test
- DORA WA subtest and the Woodcock Word Attack

High levels of correlation help to demonstrate criterion validity of the LGL Reading Assessment.

- Correlation (HFW & SORT): r=0.95 SE=0.073 (n=21)
- Correlation (WR & WI-W): r=0.92 SE=0.088 (n=21)
- Correlation (WA & W-WA): r=0.91 SE=0.097 (n=21)


LGL SP subtest and the WRAT LGL SR subtest and the Gray Oral Reading Test

Correlation (SP &WRAT): r=0.85 SE=0.210 (n=21) Correlation (SR &GORT):* r=0.65 SE=0.250 (n=21) Medium to high correlations demonstrate concurrent validity of the LGL Reading Assessment. *Lower correlations to the GORT are attributed to a high variability observed in the GORT results. Students inconsistently tested well above their grade levels on the GORT. Subsequent SR comparisons with a more consistent paper-and-pencil assessments was recommended.
New York City Schools (2009) Correlation Study

In a second comparison to one-on-one paper-and-pencil assessments performed by literacy specialists, Let’s Go Learn again achieved high correlations with statistical significance, indicating another measure of assessment validity. This study was conducted in an ethnically diverse urban New York City school in October 2009, with a sample size of 915 students. Students were tested both using the Fountas and Pinnell Benchmark Assessment and online using Let’s Go Learn DORA assessment within a two-week time period. Due to the nature of the Fountas and Pinnell Benchmark Assessment and its focus on providing a single reading level as opposed to identifying diagnostic reading skills, a DORA composite score was used, an average of the Vocabulary and Comprehension sub-tests of DORA, for the correlation to the Fountas and Pinnell score.

The correlation of the average of the DORA Vocabulary and Comprehension sub-tests to the F&P score resulted in a correlation of 0.730, which was statistically significant ($p < .01$). This high correlation between the two assessments once again provides construct validity, particularly to the subtests of vocabulary and comprehension.

Ault-Highland Schools (2010) Hierarchical Linear Modeling

In order to examine the predictive relationship between DORA subtests and the Colorado state reading achievement testing, a hierarchical linear model was developed. The relationship between an online formative assessment program, the Diagnostic Online Reading Assessment (DORA), and a state test in reading was examined in one school district ($n = 208$) using a Two-Level Time-Varying Covariate Hierarchical Linear Growth Model. It was found that all DORA subtests were significantly and positively
related to state reading test growth. See Appendix C for the complete results of this study.

*Jersey City Schools (2013) Correlation and Regression Analysis*

In 2013, a study was conducted in Jersey City (NJ) with a sample of K-12 students ($n = 4,877$). Two analyses were conducted, a correlation study and a regression analysis. The initial correlation between the New Jersey Statewide Assessment (NJ ASK) and an average of the *DORA* Vocabulary and Comprehension sub-tests showed a high level of correlation ($r = .642$), that was statistically significant ($p < .01$).

Further regression analysis examined the predictability of *DORA* subtests to the NJ ASK. After accounting for gender, ethnicity, and SES, DORA Comprehension scores are a significant predictor of NJ ASK ELA achievement ($F = 3256.228, df = 1, 4872, p < .001$). After accounting for control variables as well as DORA Comprehension, Vocabulary scores are also a significant predictor of NJ ASK ELA achievement ($F = 96.294, df = 1, 4871, p < .001$).

In summary, the final model does account for a statistically significant portion of the variance in NJ ASK ELA achievement scores. DORA Comprehension and DORA Vocabulary are significant predictors of and contribute significantly to a student’s overall NJ ASK ELA achievement. The model indicates that there is a strong connection between success on the DORA subtests of Comprehension and Vocabulary and a student’s overall scaled score performance on the ELA test of the NJ ASK. For full analysis results, please see Appendix D of this document.
DORA Internal-Consistency Reliability Study

Introduction

According to Allen & Yen (1979), reliability can be defined as the consistency between the observed scores on an assessment and the true scores. There are multiple methods of assessing the reliability of an assessment. One way is the test-retest method, comparing the consistency between one administration and a subsequent second administration of the same assessment. Another method of determining reliability is by using internal consistency, a measure of the consistency of results of items within one test. This report will present the internal consistency results for Diagnostic Online Reading Assessment (DORA) (Let’s Go Learn, Inc.).

Methods

Let’s Go Learn, Inc., a developer of online diagnostic reading and math assessments, has created an online diagnostic assessment to evaluate students’ reading abilities, Diagnostic Online Reading Assessment (DORA). DORA assesses students across seven different subskills of reading: High Frequency Words, Word Recognition, Phonics, Phonemic Awareness, Oral Vocabulary, Spelling, and Comprehension. Further, the assessment is built on an adaptive-logic platform, limiting the number of items seen by each student by both determined start point and by early termination. For each sub-test except Comprehension, there is one version of the assessment; the comprehension sub-test cycles through three versions (A, B, & C).
In order to evaluate the reliability of DORA, the Rasch model was used to evaluate overall model fit for each sub-test. A sample of 17,856 students in grades K-12 taken from schools nationwide (6 districts from California, Colorado, Hawaii, and Virginia) was used.

Analysis

Analysis showed overall unidimensionality of each sub-test as well as good overall model fit. Reliability levels were strong (see Table 1).

Table 1

<table>
<thead>
<tr>
<th>Sub-test</th>
<th>Reliability (α)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Frequency Words</td>
<td>.74</td>
</tr>
<tr>
<td>Word Recognition</td>
<td>.89</td>
</tr>
<tr>
<td>Phonics</td>
<td>.75</td>
</tr>
<tr>
<td>Phonemic Awareness</td>
<td>.45*</td>
</tr>
<tr>
<td>Oral Vocabulary</td>
<td>.94</td>
</tr>
<tr>
<td>Spelling</td>
<td>.96</td>
</tr>
<tr>
<td>Comprehension</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>.87**</td>
</tr>
<tr>
<td>B</td>
<td>.92***</td>
</tr>
<tr>
<td>C</td>
<td>.89****</td>
</tr>
</tbody>
</table>

* n = 4878
** n = 3420
*** n = 10534
**** n = 1596

Conclusions

For most high-stakes assessments, a reliability of .7 or higher is considered a reasonable level for strong reliability. Since DORA is a formative assessment, the levels indicated for all sub-tests, with the exception of Phonemic Awareness, more than demonstrate a high level of reliability. The discrepancy on the reliability score for the
Phonemic Awareness sub-test can be attributed to a number of factors. First of all, since Phonemic Awareness is only administered to K-2 students or any student meeting certain adaptive criteria in grades 3-12, the sample size for the Phonemic Awareness analysis was much smaller. In addition, the Phonemic Awareness subset is designed as a screener and consists of only 9 items, a much smaller sub-test than any of the others; it is not unexpected to see reliability levels decrease as the number of items assessed decreases. Furthermore, in this type of assessment, there is the issue of missing data, which will affect the reliability levels. Since no one student ever takes all test items, there is always missing data. The missing data there serves the restrict variability even further, in turn increasing the standard error as well as the mean square error.
Introduction

According to Allen & Yen (1979), reliability can be defined as the consistency between the observed scores on an assessment and the true scores. There are multiple methods of assessing the reliability of an assessment. One way is the test-retest method, comparing the consistency between one administration and a subsequent second administration of the same assessment. Another method of determining reliability is by using internal consistency, a measure of the consistency of results of items within one test. This report will present the internal consistency results for Diagnostic Online Reading Assessment Spanish (DORA Spanish) (Let’s Go Learn, Inc.).

Methods

Let’s Go Learn, Inc., a developer of online diagnostic reading and math assessments, has created an online diagnostic assessment to evaluate Spanish-speaking students’ reading abilities, Diagnostic Online Reading Assessment Spanish (DORA Spanish). DORA Spanish assesses students across seven different subskills of reading: High Frequency Words, Word Recognition, Phonics, Phonemic Awareness, Oral Vocabulary, Spelling, and Comprehension. Further, the assessment is built on an adaptive-logic platform, limiting the number of items seen by each student by both determined start point and by early termination.
In order to evaluate the reliability of *DORA Spanish*, the Rasch model was used to evaluate overall model fit for each sub-test. A sample of 6,244 students in grades K-12 taken from schools nationwide was used.

*Analysis*

Analysis showed overall unidimensionality of each sub-test as well as good overall model fit. Reliability levels were strong (see Table 1).

**Table 1**

*DORA Sub-test Reliability (α)*

<table>
<thead>
<tr>
<th>Sub-test</th>
<th>Reliability (α)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Frequency Words</td>
<td>.94</td>
</tr>
<tr>
<td>Word Recognition</td>
<td>.94</td>
</tr>
<tr>
<td>Phonics</td>
<td>.77</td>
</tr>
<tr>
<td>Phonemic Awareness</td>
<td>.44</td>
</tr>
<tr>
<td>Word Meaning</td>
<td>.61</td>
</tr>
<tr>
<td>Spelling</td>
<td>.86</td>
</tr>
<tr>
<td>Comprehension</td>
<td>.84</td>
</tr>
</tbody>
</table>

\[n = 6,244\]

*Conclusions*

For most high-stakes assessments, a reliability of .7 or higher is considered a reasonable level for strong reliability. Since *DORA Spanish* is a formative assessment, without high-stakes outcomes, the levels indicated for all sub-tests, with the exception of Phonemic Awareness, more than demonstrate a high level of reliability. The discrepancy on the reliability score for the Phonemic Awareness sub-test can be attributed to a number of factors. First of all, since Phonemic Awareness is only administered to K-2 students or any student meeting certain adaptive criteria in grades 3-12, the sample size for the Phonemic Awareness analysis was much smaller. In addition, the Phonemic
Awareness subset is designed as a screener and consists of only 9 items, a much smaller sub-test than any of the others; it is not unexpected to see reliability levels decrease as the number of items assessed decreases. Furthermore, in this type of assessment, there is the issue of missing data, which will affect the reliability levels. Since no one student ever takes all test items, there is always missing data. The missing data there serves the restrict variability even further, in turn increasing the standard error as well as the mean square error.
Test-Retest is the ability of a test to be taken once and then immediately again and have similar results, offering another measure of test reliability. Let’s Go Learn undertook a second test-retest study in Q3 of 2003. This data was added to an earlier study performed in Q1 of 2003 in order to provide a larger sampling and to reduce the margin of error. The combined results were once again excellent. Variability was low, meaning that the LGL Reading Assessment is very precise and can be re-administered with low bias. Sample size: n=225

Grade level deltas are calculated by first taking the absolute value of the change (or delta) between each test and retest subtest score. Next the mean is calculated from these deltas. All scores are in grade levels, so a grade level delta of 0.5 means one half of a grade level. SE is standard error for mean delta in grade levels.

- **High-Frequency Words** - Grade level delta: 0.39 SE=0.11
- **Word Recognition** - Grade level delta: 0.19 SE=0.12
- **Word Analysis** - Grade level delta: 0.16 SE=0.07
- **Word Meaning** - Grade level delta: 0.60 SE=0.19
- **Spelling** - Grade level delta: 0.27 SE=0.10
- **Silent Reading** - Grade level delta: 0.35 SE=0.13
References


Appendix A: Silent Reading Leveled Passages Design Specifications

Fact v. Fiction

The Let’s Go Learn silent reading passages are all non-fiction passages, an intentional design choice. When students are taught to read in the primary grades, they are often taught using primarily fictional passages, in which a more predictable story structure and vocabulary can support them as they master word identification skills. Roughly at the beginning of fourth grade, instruction begins to focus more heavily on reading for the sake of learning new factual information. Often this transition takes place quite suddenly when students are asked to read content area textbooks in addition to stories. Many of the self-extending strategies that children develop for comprehending narrative text are less useful in approaching non-fiction, and additional new strategies are needed to help them make sense of non-fiction passages. Yet students often do not receive explicit instruction in how to read non-fiction. As a result, reading performance may show a sudden, marked “decline” at the point of transition to non-fiction texts, when in fact what has significantly and abruptly changed is the nature of the task demands.

DORA is structured to avoid this trap by using only factual passages and no fiction. The reasons for choosing non-fiction over fiction, even at the lower grade levels, are twofold. First, we believe that children should have opportunities to encounter non-fiction (and to learn the relevant strategies for unpacking it) from early on in school; we reject the simplistic notion that “reading to learn” should begin only at the intermediate grade levels.

Our second reason for selecting non-fiction over fiction passages is that assessment results for non-fiction passages are generally a more conservative measure of
comprehension. While a student who excels in reading fiction may have much greater
difficulty with non-fiction passages, it is far more rare to encounter a student whose non-
fiction reading level is higher than her/his fiction reading level. Assessing a child on a
fiction passage could result in a misleadingly high score that masks areas where the child
would benefit from additional opportunities and support for her/his reading development.
By using non-fiction passages, Let’s Go Learn is able to provide a score that – while it
may be lower than what the child could do on fiction passages – offers diagnostic
information about the child’s reading ability on the type of material that he or she may
encounter most in further schooling and in “real life.”

Topics

The DORA silent reading passages draw on material set out in major content
standards from different states. Passages for each grade level are based on selected topics
from the social studies, fine arts, and science standards for that grade level. The topics
were chosen for multiple reasons. First, the topics were narrowed from a list of possible
choices to those likely to be of greatest interest to students near that grade level. Second,
the topics were chosen that would most likely be part of all schools’ curriculum for that
grade level; obscure topics that might not be addressed in detail were eliminated. Finally,
topics were chosen that were not so obscure that children near that grade level would be
unlikely to have sufficient background knowledge to be able to make sense of the
passages.
Stylistic Design

Passages are written in a journalistic format. The passages stylistically most resemble a newspaper or encyclopedia article in format. They begin with an introductory paragraph that presents the main idea of the passage, often in a manner designed to get the reader’s attention--for example, via a leading question or a reference to common phrases or occurrences.

The rest of each passage is written in organized, sequential paragraphs. There is not a concluding summative paragraph on most passages as such paragraphs are not typical in this genre. The style is designed to be as similar to textbook or authentic informational writing as possible.

Comprehension Questions

The DORA comprehension sub-test provides information both on students’ knowledge acquisition through what they read and on students’ ability to analyze and apply that knowledge in a variety of ways. The questions are divided into two main categories: factual questions and higher-level, extending questions. Questions that are factual require the student to report on information that was explicitly stated in the text. Extending questions require students to take information from the text and apply it--to their own lives, to other things they know about the world, or in order to draw conclusions.
In most classrooms, the majority of questions asked about a reading passage are factual. This may be adequate for assessing simple comprehension of sentences, but in order to assess students’ deeper comprehension of text, higher-level questioning strategies must be employed.¹ *DORA* looks at both dimensions in order to provide a more complete assessment of a student’s comprehension.

Each passage is followed by 6 questions. Three of these are factual questions, which assess knowledge acquisition on a basic level—can the student read and obtain information from an informational passage? There are also three extending questions.

One of these extending questions requires the student to define a vocabulary word from the text. Since words are selected that are unlikely to be within most children’s active or passive vocabulary at the targeted grade level, these questions assess a student’s ability to find a meaning for an unfamiliar word in the context of the passage. This is considered an extending question because it not only requires students to use knowledge from the text, but it also requires them to draw on previous word knowledge in order to draw conclusions.

The final two questions are extending questions that require students to do one of several tasks. They might ask a student to answer “how” or “why” questions about facts in the text; they might ask students to apply the facts to their own lives or experiences; or they might ask students to think of similar situations to those described in the text.

All *DORA* comprehension questions are designed to assess the students’ understanding of the passage in a way that does not confound their comprehension with their knowledge of language arts terminology such as “main idea” or “simile.”

¹ For more information about authentic questioning strategies, please reference Robert Ruddell’s *Becoming an Influential Teacher.*
Assessments that rely on such terms are substantially dependent on students’ understandings of the terms themselves and thus function more as vocabulary assessments than as assessments of comprehension. A student may be able to identify the main idea of a passage without knowing that it is called a “main idea.” Or a child may be able to summarize brilliantly when asked “What was this passage about?” but be shut down and flummoxed by a question that asks for a “summary.” By avoiding questions that rely on knowledge of language arts jargon, DORA questions reduce inaccurate comprehension scores that are likely when the student understands a passage well but does not know the terminology. We minimize the use of jargon in order to provide a less filtered, purer assessment of the underlying comprehension.
Appendix B: Grade Level Justifications and Parallels for DORA Silent

Reading Leveled Passages

The Let’s Go Learn silent reading passages (versions A-C as of 9/2007) are evaluated across several different categories to ensure that the results of testing remain the same, regardless of which passage is used. Attached is a spreadsheet detailing exact data ranges for all categories that are considered during passage leveling.

Passage grade leveling is dependent on several different factors, including passage length, sentence length, sentence complexity, and vocabulary. There are several different measures and standards available for leveling. Let’s Go Learn employs the Flesh-Kincaid leveling system, modified with considerations for sentence complexity and vocabulary. If a leveling system such as Flesh-Kincaid is used exclusively, it does not take into account whether the level determined is based more on vocabulary or on sentence complexity (as F-K levels use both). A fourth grade passage that is level 4.0 because of vocabulary is very different from a passage that is level 4.0 because of sentence structure. By using F-K to obtain a numerical value and then aligning passages to Let’s Go Learn standards using a variety of data points, it is assured that the passages are consistent in terms of both vocabulary and structure across each grade level and from one grade level to the next. The data point score ranges are kept narrow, and further passages are leveled into the already-established score ranges.
**Passage Length and Paragraphing**

The length of *DORA* silent reading passages increases steadily as the level increases. First grade passages begin at 100-125 words per passage (wpp) and increase by 25-50 wpp per grade, to as many as 395-415 wpp at the 12th grade level. By not dipping below 100 wpp, even at the first grade level, the passage is long enough to provide detailed miscue analysis. Another concern for length is test exhaustion: passages more than 500 words long could lead to assessment data reflecting test exhaustion rather than ability (especially if a student must read multiple passages).

Paragraph length also expands because of increases in sentence length and complexity. The number of sentences per passage (spp) also increases slowly by grade level (12-20 at first grade; 28-30 by seventh grade), until the seventh grade level, at which time the rate of spp begins to decrease as the complexity of sentences markedly increases.

**Sentence Length and Complexity**

The length of sentences per passage (spp) increases steadily as the grade level increases. At the first grade level, the average sentence length ranges from 6.5 – 8.5 words per sentence (wps). By the seventh grade level, the average has increased to 12.1-14.7 wps. At the twelfth grade level, the average sentence length has increased to 20.8-21.8 wps.

Along with sentence length, sentence complexity is also considered. Data is taken for each passage to show the percentage of simple, compound, complex, and compound-complex sentences. The percentage of simple sentences per passage declines steadily as the percentage of other varieties of sentences increases. While compound sentences are
slightly harder to read, complex sentences increase difficulty even more as they complicate independent clauses with a dependent clause variable.

At the first grade level, the percentage of simple sentences ranges from 82-85%, with only 5-12% compound sentences and 2-12% complex sentences. There are no compound-complex sentences at the first grade level. By the seventh grade level, the percentage of simple sentences has decreased to 64-79%, with 0-15% compound sentences and 7-32% complex sentences. At the seventh grade level 0-4% of the sentences may be compound-complex. Finally, by the twelfth grade level, the range of simple sentences is 21-68% (a slightly larger range, due to the overall complexity of writing at this level), with 5-32% compound sentences and 16-37% complex sentences. At this level there may be as many as 5-16% compound-complex sentences. As evidenced by these examples, the overall range of simple sentences declines substantially while the range of compound sentences increases slightly, and the range of complex and compound-complex sentences increases noticeably.

Other considerations in the complexity of sentence structure are the rate of prepositional phrases per sentence and the rate of verbals (gerunds, infinitives, and participles) per sentence. Prepositional phrases increase the difficulty of a sentence by adding information that could have been presented in its own independent clause as a phrase attached to an independent clause. The rate of prepositional phrases per sentences increases steadily by grade level as well. At the first grade level the average number of prepositional phrases per sentence is 0.2-0.6 per sentence. By seventh grade that rate has increased to 0.85-1.9 prepositional phrases per sentence. Finally, by the twelfth grade,
the average increases to 2.2-2.6 prepositional phrases per sentence, an overall increase of 2 phrases per sentence from the first grade.

Verbals increase sentence difficulty by presenting a verb in the position of a noun or adjective. This often confuses students—who are used to seeing these words as the actions in the sentence—and may create comprehension problems for those students who actively use syntactical cues when reading. The rate of verbals per sentence also increases gradually (as verbals in general are not used with the frequency of other grammatical structures, like prepositional phrases). At the first grade level, the rate of verbals per sentence is very low, only 0.05-0.25 verbals per sentence. At the seventh grade level, the average has risen slightly to 0.5-0.6 verbals per sentence. By the twelfth grade the average is 0.58-0.84 verbals per sentence.

**Vocabulary**

One large concern when leveling passages is the complexity of vocabulary. This becomes even more of an issue when writing non-fiction passages because of course, there are terms that must be introduced that are specific to each topic. These elements of specialized vocabulary can complicate a leveling system, as they often increase the reading level dramatically. The LGL silent reading passages continue to use terminology and specialized vocabulary in the leveled passages but complement these more complicated, unfamiliar terms with other, easier vocabulary, thus keeping the leveling patterns consistent for each grade level, regardless of the specialized vocabulary.

That being said, the difficulty of vocabulary overall can be understood by looking at the average number of syllables per word (spw). This range increases steadily by
grade level also. At the first grade level, there are few polysyllabic words (only 10-20%), and the average number of syllables is only 1.1-1.2 spw. By the seventh grade, the number of polysyllabic words has increased (now 31-35%), and the average number of syllables has increased to 1.4-1.5 spw. Finally, by the twelfth grade level, polysyllabic words are commonplace (35-45%), and the average number of syllables is 1.56-1.73 spw.

The introduction of polysyllabic words of more than 3 syllables is also an indication of vocabulary complexity. Two-syllable words only increase in frequency until roughly the sixth grade level, at which point the rate becomes more consistent as the rate of polysyllabic words of 3 syllables or more increases. Four-syllable and 5-syllable words are not introduced in the text with any notable frequency until between grades 5 and 7. At that time, their use is minimal and increases only slightly until the twelfth grade (when their rates are 3-6% and 1-2%, respectively). In general, as the grade level of the passage increases, the number of single-syllable words decreases, and the number of more complex polysyllabic words increases.
Appendix C: The Relationship between Online Formative Assessment Scores and State Test Scores Using Multilevel Growth Modeling

Running head: RELATIONSHIP BETWEEN ONLINE FORMATIVE ASSESSMENT

The Relationship between Online Formative Assessment Scores and State Test Scores Using Multilevel Growth Modeling

Aryn C. Karpinski, Ph.D.
Kent State University

Jerome V. D’Agostino, Ph.D.
The Ohio State University

Anne-Evan K. Williams, M.A.
The Ohio State University

Sue Ann Highland
Weld County RE-9 (Ault-Highland)

Author Note

Correspondence concerning this paper should be addressed to Aryn C. Karpinski, Ph.D., Department of Foundations, Leadership, and Administration, Evaluation and Measurement, Kent State University, 316 D White Hall, Kent, Ohio 44240. E-mail: akarpins@kent.edu
Abstract

As technology-based teaching and assessment is gradually used to support and/or replace traditional forms of evaluation, the need to examine the extent to which these methods are educationally sound is in high demand. The relationship between an online formative assessment program, The Diagnostic Online Reading Assessment (DORA), and a state test in reading was examined in one school district (N = 208) using a Two-Level Time-Varying Covariate Hierarchical Linear Growth Model. It was found that all DORA subtests were significantly and positively related to state reading test growth. Information presented in this study can provide practical implications to district-wide implementation of supplemental reading instruction in an online environment.
The Relationship between Online Formative Assessment Scores and State Test Scores Using Multilevel Growth Modeling

The main goal of the current study was to examine the relationship between online formative assessments and summative, yearly state proficiency test scores. Specifically, the relationship between one online formative assessment program in reading, known as the Diagnostic Online Reading Assessment (DORA), and state test scores in reading (i.e., the Colorado Student Assessment Program [CSAP]) was examined in four cohorts across elementary, middle, and high school in beginning in the 2004/2005 academic year and ending in 2009/2010. This investigation used Hierarchical Linear Growth Modeling (HLGM) to address the following research question: (1) What is the relationship between online formative assessment score growth and state test score growth?

The purposes in conducting this study include the following: (1) To support the burgeoning literature outlining the role of the Internet, and technology in general, in teaching and learning, and (2) To bolster support for federal initiatives and administrative demands for more efficient ways to meet state standards. As technology-based teaching and assessment is gradually used to support and/or replace traditional forms of evaluation, the need to examine the extent to which these methods are educationally sound is in high demand. Overall, information presented in this study can provide practical implications to district-wide implementation of supplemental reading instruction in an online environment.

Literature Review

E-learning (i.e., learning that is facilitated by electronic technologies) is referred to as part of the equipment of 21st Century scholarship (Buzzetto-More & Guy, 2006). However, e-learning is only half of the equation as government mandates have required schools to use data to inform decision making. The use of data has necessitated the development of improved information technology and access to computers and high-speed Internet in schools (Petrides, 2006). Thus, the other half of the equation is the use of data rendered from e-learning, or e-assessment, which entails using electronic technologies to drive student learning and assessment as with formative assessment (Ridgway, McCusker, & Pead, 2004).

Formative assessment can be briefly defined as the use of diagnostic formal and informal assessments to provide feedback to teachers and students over the course of instruction for the purpose of improving performance and achievement (Boston, 2002). Previous research in this area has primarily focused on traditional formative assessment practices (e.g., paper-and-pencil quizzes), with the current literature beginning to examine the effectiveness of computerized or Internet-based, automated formative assessment programs. The overall consensus from the traditional body of literature is that formative assessment is an essential component of classroom procedure, and that its proper use can raise standards and achievement (Black & William, 1998a), with the latest studies of technology-based formative assessment beginning to echo these findings.
Methods

Existing DORA data was provided from one school district in Colorado from an online testing company, and existing CSAP data was provided from the same school district by the Colorado Department of Education (CDE). Data were linked anonymously producing four cohorts of students across grades 3 through 11.

It is hypothesized that student formative assessment score growth will be significantly and positively related to student state test score growth. The analytical method used was a Two-Level Time-Varying Covariate HLGM (Singer & Willett, 2003). Time (i.e., measured in months with one unit being every 3 to 6 months) and DORA scores (i.e., the time-varying covariate) were used in Level 1. Demographic covariates such as gender, ethnicity, SES, and ESL/ELL status were incorporated into Level 2 of the model. CSAP scores were used as the outcome variable. Several models were run, using four DORA subtests (i.e., Word Recognition [WR], Oral Vocabulary [OV], Spelling [SP], and Reading Comprehension [RC]) as the time-varying covariate in separate models. These subtests did not have ceiling effects compared to the other subtests.

The model at Level 1 was the following:

\[ Y_{it} = \pi_{0t} + \pi_{1t}(DORA)_{it} + \pi_{2t}(Time)_{it} + \epsilon_{it} \]

where \( Y_{it} \) is the student’s CSAP score for time \( t \) for student \( i \), \((Time)_{it}\) is the elapsed years/months since DORA implementation, \( (DORA)_{it} \) is the time-varying predictor for a student at a given time point, \( \pi_{0t} \) (i.e., the intercept) is a student’s initial CSAP score, \( \pi_{1t} \) is the linear growth coefficient (i.e., for DORA) along with \( \pi_{2t} \) (i.e., the growth rate over all years/months), which represents the child’s expected change in CSAP score for unit year/month increase. This individual growth-curve model assumes that \( \epsilon_{it} \) is individual student error.

The model at Level 2 was the following:

\[ \pi_{0i} = \beta_{00} + \beta_{01}(SEX)_{i} + \beta_{02}(ETHNIC)_{i} + \beta_{03}(ESLELL)_{i} + \beta_{04}(FREERED)_{i} + r_{0i} \]

\[ \pi_{1i} = \beta_{10} \]

\[ \pi_{2i} = \beta_{20} \]

where \( \pi_{0i}, \pi_{1i}, \) and \( \pi_{2i} \) are the individual-specific CSAP score parameters (i.e., initial status, DORA growth, and growth rate), \( \beta_{00} \) is the baseline expectation (i.e., initial CSAP status) for the demographic predictors coded as 0, \( \beta_{10} \) is the expected change of the CSAP controlling for the DORA time-varying covariate, and \( \beta_{20} \) is the expected linear change of the CSAP for the demographic predictors coded as 0. Finally, \( r_{0i} \) is a residual.

The demographics were not modeled for all the intercepts and slopes in the Level 2 equations. The final estimates of the Level 2 variance components for the demographic variables in the intercept and slope equations (i.e., the ones removed above) were very small with most being significant. When this occurs, it is common to fix some of the effects (i.e., eliminate the error term). Modeling small error variances can be problematic due to the increase in parameters.
estimated and loss of degrees of freedom. The most parsimonious model necessitates fixing these effects. Thus, for the two equations above not including the demographics, the effects were set as fixed (i.e., the error term was eliminated) for every final model.

**Data Sources/Evidence:**

*Data.* Existing DORA and CSAP data were used (*N = 208*) from 2004/2005 to 2009/2010 for students in grades 3 through 11. For the CSAP, students’ test scores from the spring of 2005 were the first data point. Two state reading test scores were obtained before DORA implementation and three afterwards. For DORA, students’ test scores from fall of 2006 represented the first data point. Students were tested approximately three times during the school year, with a possible total of 11 DORA assessments for the years investigated in the current study. See Table 1 for the final analysis sample demographic information.

*Measures.* DORA tests across seven subtests: (1) High-Frequency Words, Phonics, Phonemic Awareness, Word Recognition, Oral Vocabulary, Spelling, and Reading Comprehension. With the exception of the Spelling subtest which is a generative test, all test items are multiple-choice. DORA results are returned as grade-level equivalencies.

The CSAP is administered each spring, yielding a single, scaled score (i.e., reading score in the current study) for each student each year. The state scores were based on a scale that ranged from 0 to 1000 depending on the grade level assessed, with a score of approximately 550 as the cut-point for proficiency each year. At each grade level, students are assessed using 40 to 70 multiple choice items depending on the grade level, developed to assess student knowledge of grade-level indicators, identified as the Colorado Model Content Standards for that particular grade level. The tests across grades were vertically equated.

**Results/Conclusions:**

*HGLM Results.* The full model-building strategy was implemented for each DORA subtest as the time-varying covariate (i.e., One-way Random Effects ANOVA, Unconditional Model, Conditional Growth Model, Full Model). Only the Full Models will be discussed below; however, the specificity of the models above warrant a more detailed discussion (e.g., model comparisons including random effects, model deviance comparisons). The findings pertaining to the covariation between DORA and CSAP scores and DORA growth rates over time are presented in Table 2. Each column provides the results from one HGLM analysis (i.e., one per DORA subtest).

As can be seen, all initial DORA average scores across the four subtests were significantly greater than zero. The growth rates for all subtests were significantly greater than zero. The DORA and CSAP covariation results indicated that students gain in DORA over time covaried positively and significantly with their CSAP gain on all DORA subtests. For every one point increase in WR score, there was a 1.78 point increase in the state test score, and for every one point increase in OV score, there was a 2.17 point increase in the state test score. For every one point increase in SP score, there was a 3.41 point increase in the state test score, and for every one point increase in RC score, there was a 2.75 point increase in the state test score.
This is an interesting finding as typically the RC subtest is viewed as the most similar in structure and content to state reading tests (Let’s Go Learn, Inc. ©, 2009). The RC subtest attempts to provide a window into the semantic domain of a learner’s reading abilities. Children silently read passages of increasing difficulty and answer questions about each passage immediately after they read it. The questions for each passage are broken up into three factual questions, two inferential questions, and one contextual vocabulary question. This is typically how many state reading tests structure their exams.

As indicated above, the SP subtest surprisingly was related to the fastest CSAP growth rates in students. SP is a generative process as opposed to a decoding or meaning-making process as seen in most assessments of reading comprehension, which does not support the finding as noted above. Additionally, it is natural for young readers’ spelling abilities to lag a few months behind their reading comprehension abilities (Bear, Invernizzi, Templeton, & Johnston, 2000).

**Educational/Scientific Importance of Study:**

As mentioned before, the goal in this study was to examine if student CSAP growth is related to student DORA growth. The hypothesis was that student CSAP growth will be significantly and positively related to student DORA growth. Thus, the relationship of interest in addressing this research question is with the time-varying covariate and the state reading test. The hypothesis was supported in the Full Models controlling for the demographics in that the DORA scores for all the subtests were positively and significantly related to state reading test scores, indicating that these subtests are demonstrating a correlated growth in students reading to the state testing regardless of demographic status.

Although causal inference is limited, the findings from this study have implications in a number of contexts and for many groups of people. Overall, the demonstrated relationship can provide teachers and administrators with evidence to warrant the continued use of technology-based formative assessment practices. Specific to the results, these subtests are further supported as a learning tool to gauge, or perhaps predict, student performance on the reading state test. This relationship is reassuring given the number of educators who are using technology-based teaching methods, and the number of administrators who are seeking to increase the use of technology as a learning tool in their schools.

For administrators, the demand for school systems, individual schools, and teachers to be accountable for student performance has increased considerably over the past two decades. This demand for accountability relates to a direct measurement of attainment of educational standards and objectives. Higher student numbers in schools and colleges can pressure administrators to find alternative, more efficient routes to raise test scores and meet federal and state demands. Thus, the positive relationship between online formative assessment and student state test scores can render support for school districts and individual schools to obtain grant money and permission needed to acquire the necessary online formative assessment programs to alleviate some of the strain. Overall, the results from this study support the use of DORA as a way to measure and attain various educational standards, such as having students pass and excel on the end-of-year, summative state exam.
The major methodological limitation in the current research question was the lack of a control group. Although the usefulness of studies such as the current investigation and related research questions have a place in the research process (e.g., theory generation, determining correlates for future experimental research), the conclusions drawn cannot extend past what the design and analyses can demonstrate. The findings in the current research question cannot state that computerized/online formative assessment use causes increased state test scores, only that DORA and CSAP growth are positively related. Future research should consider implementing a similar design, but also obtain an adequate control group (e.g., other districts who are not using DORA).

Another major methodological limitation includes the use of one school district, which is a considerable threat to external validity. Sampling a number of school districts with multiple elementary, middle, and high schools could have added to the validity of the results. With regards to the DORA and CSAP data, the DORA data are from 2006/2007 to current and the CSAP data range from 2004/2005 to present as well. Moreover, there are more DORA time points than CSAP time points. Generally, HLM can accommodate time-unstructured data such as the above; however, the accuracy and validity of results (e.g., statistical conclusion validity) many times depends on how closely the data are measured (i.e., same time/day compared to several days/weeks apart) in a time-varying covariate model (Biesanz, Deeb-Sossa, Papadakis, Bollen, & Curran, 2004).

With regards to the HLM assumptions, some violations were noted in examining linearity, normality, and homogeneity. Violations of one assumption can also complicate the results or interpretation of other assumptions. Future studies should use multiple school districts with more data collection time points to have more flexibility to eliminate outliers and evaluate assumptions more accurately.
Selected References


### Table 1

*Demographic Information of the Final Analysis Sample (N = 208) for Grades 3 through 11 across the 2004/2005 to 2009/2010 Academic Years*

<table>
<thead>
<tr>
<th>Demographic Information</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cohort</strong></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>47 (22.6)</td>
</tr>
<tr>
<td>2</td>
<td>52 (25.0)</td>
</tr>
<tr>
<td>3</td>
<td>48 (23.1)</td>
</tr>
<tr>
<td>4</td>
<td>61 (29.3)</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>104 (50.0)</td>
</tr>
<tr>
<td>Female</td>
<td>104 (50.0)</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
</tr>
<tr>
<td>White (Non-Hispanic)</td>
<td>142 (68.3)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>60 (28.8)</td>
</tr>
<tr>
<td>Black (Non-Hispanic)</td>
<td>3 (1.4)</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>2 (1.0)</td>
</tr>
<tr>
<td>American Indian/Alaskan Native</td>
<td>1 (.5)</td>
</tr>
<tr>
<td><strong>Free/Reduced Lunch</strong></td>
<td></td>
</tr>
<tr>
<td>Eligible</td>
<td>93 (44.7)</td>
</tr>
<tr>
<td>Not Eligible</td>
<td>115 (55.3)</td>
</tr>
<tr>
<td><strong>English Language Learner (ESL/ELL)</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>26 (12.5)</td>
</tr>
<tr>
<td>No</td>
<td>182 (87.5)</td>
</tr>
<tr>
<td><strong>Individualized Education Program (IEP)</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>-</td>
</tr>
<tr>
<td>No</td>
<td>208 (100.0)</td>
</tr>
</tbody>
</table>
Table 2

**HLGM Models for CSAP and Time Predicting DORA Subtests**

<table>
<thead>
<tr>
<th>Fixed Effects</th>
<th>WR</th>
<th>OV</th>
<th>SP</th>
<th>RC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model for Initial CSAP Status (π₀)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept (β₀₀)</td>
<td>633.10 (4.58)**</td>
<td>633.16 (4.50)**</td>
<td>635.45 (4.42)**</td>
<td>634.63 (4.32)**</td>
</tr>
<tr>
<td>Sex (β₀₁)</td>
<td>7.02 (5.45)</td>
<td>7.45 (5.22)</td>
<td>6.98 (4.99)</td>
<td>7.83 (4.91)</td>
</tr>
<tr>
<td>Ethnicity (β₀₂)</td>
<td>-2.51 (6.60)</td>
<td>-6.6 (6.21)</td>
<td>-4.70 (6.16)</td>
<td>-.43 (5.82)</td>
</tr>
<tr>
<td>ESL/ELL (β₀₃)</td>
<td>-37.86 (8.98)**</td>
<td>-35.69 (8.48)**</td>
<td>-35.53 (8.52)**</td>
<td>-32.43 (7.63)**</td>
</tr>
<tr>
<td>Free/Reduced Lunch (β₀₄)</td>
<td>-15.03 (5.74)*</td>
<td>-14.58 (5.41)**</td>
<td>-12.98 (5.26)*</td>
<td>-13.56 (5.18)*</td>
</tr>
<tr>
<td>Model for DORA Growth Rate (π₁)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept (β₁₀)</td>
<td>1.78 (.52)**</td>
<td>2.17 (.58)**</td>
<td>3.41 (.67)**</td>
<td>2.75 (.46)**</td>
</tr>
<tr>
<td>Model for CSAP Growth Rate (π₂)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept (β₂₀)</td>
<td>12.49 (.87)**</td>
<td>11.65 (.92)**</td>
<td>10.69 (.94)**</td>
<td>10.02 (.97)**</td>
</tr>
</tbody>
</table>

*Note. *p < .05; **p < .01; ***p < .001*
Appendix D: The predictive values of DORA scores in New Jersey Statewide Assessment (NJ ASK)

Study Design:

In order to examine the predictive value of the subtest scores of Let’s Go Learn’s Diagnostic Online Reading Assessment (DORA) on the English/Language Arts scaled scores of the New Jersey Statewide Assessment (NJ ASK), a regression analysis was performed. Data was gathered district-wide from the Jersey City Public Schools, with a final sample size of 4,877 students. DORA scores were gathered in the spring of 2013, and NJ ASK scores were obtained also during 2013.

Analysis Summary:

Previous DORA research has suggested that the subtests of Comprehension and Vocabulary are often more strongly predictive of state testing scores in English/Language Arts (Karpinski, 2010). For this reason, and because these two subtests were of greatest interest to the study, these were the two independent variables included in the analysis.

In addition to the two variables of study interest, variables for gender, ethnicity, and socio-economic status were also included as control variables. Gender was coded 1 for male, 0 for female; ethnicity was coded 1 for minority status, 0 for white/Caucasian; and socio-economic status was coded 1 for free and/or reduced lunch eligibility, 0 otherwise.

The dependent variable in this study is the student scaled score from the 2013 NJ ASK English/Language Arts (ELA) test.

Analysis and Conclusions

The means and standard deviations of both the independent and dependent variable are shown in Table 1.

Table 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>NJ ASK ELA</td>
<td>197.05</td>
<td>26.610</td>
</tr>
<tr>
<td>Gender</td>
<td>.50</td>
<td>.500</td>
</tr>
</tbody>
</table>
Ethnicity  .91    .293
SES          .80    .398
Comprehension (CO)  5.196    3.782
Vocabulary (VO) 5.263    2.018

Note.  \( N = 4877 \)

The model for this analysis was derived using a forced-entry syntax with a forward selection logic. Control variables were entered first as a block, then study variables were entered by their highest correlation with the dependent variable: first Comprehension (.638), followed by Vocabulary (.502).

The correlation matrix was also examined to look for evidence of possible multicollinearity. There was, indeed a statistically significant correlation between both of the independent variables of study interest, Vocabulary and Comprehension. Further examination of the collinearity diagnostics indicate that none of the variables had a tolerance of less than .25, nor a condition index of greater than 30, while there is not variable loading on more than one dimension, suggesting that multicollinearity is not a problem for this model.

Further, the scatterplot between observed and predicted math scores appears to be both linear and positive, indicating a strong model and meeting the assumption of linearity between \( Y \) and \( \hat{Y} \). The assumption of homoscedasticity appears reasonable, as the plot appears evenly distributed, with no fan, funnel, or bow-tie-shaped data. From examining a histogram of the residuals, it appears that the residuals are normally distributed. The P-P Plot of the Residuals demonstrates little variance from 45°, suggesting that the assumption of normally distributed residuals appears to have been met as well.

The final model included all three control variables, as well as the two variables of study interest, with the following parameter estimates:

\[
\hat{Y} = 176.229 -5.787(\text{gender}) -5.543(\text{ethnicity}) - 1.911(\text{SES}) + 3.750(\text{CO}) + 1.877(\text{VO})
\]

The null hypothesis that the model does not provide any significant explanation is rejected at the .05 level of significance. The model is statistically significant \( (F = 746.225, df = 5, 4871, p < .001) \).

After accounting for gender, ethnicity, and SES, DORA Comprehension scores are a significant predictor of NJ ASK ELA achievement \( (F = 3256.228, df = 1, 4872, p < .001) \). After accounting for
control variables as well as DORA Comprehension, Vocabulary scores are also a significant predictor of NJ ASK ELA achievement \( (F = 96.294, df = 1, 4871, p < .001) \).

The total proportion of variance explained by the final model is \( R^2 = .433 \), indicating that 43.3% of the total variability in Jersey City NJ ASK ELA achievement is accounted for by the model. After controlling for gender, ethnicity, and SES differences in NJ ASK ELA scores within our sample, the addition of DORA Comprehension scores to the model increased the amount of variability accounted for by 38.6%, from 3.6% to 42.2%. Additionally, after controlling for differences in both DORA Comprehension scores, the addition of DORA Vocabulary scores to the model increased the amount of variability accounted for by an additional 1.1%.

The variable with the largest proportion of unique explained variance in NJ ASK ELA achievement is DORA Comprehension, with 39.6% of the variance uniquely explained by the variable. The part correlation of DORA Vocabulary scores shows that 10.6% of the variance is uniquely explained by reading scores.

The variable with the largest proportion of remaining variance, after controlling for the effects of Vocabulary and demographic controls, was also DORA Comprehension scores, with 46.6% accounted for, after partialing out the effects of the other variables. After controlling for the effects of Comprehension and demographic variables, the remaining variance accounted for by DORA Vocabulary scores was 13.9%, after partialing out the effects of the other variables. Overall, this indicates that Comprehension scores contributes more than three times that of Vocabulary scores to the variance accounted for in the model, with a small amount of additional variance accounted for by the control variables.

**Results/Conclusions:**

In summary, the final model does account for a statistically significant portion of the variance in NJ ASK ELA achievement scores. DORA Comprehension and DORA Vocabulary are significant predictors of and contribute significantly to a student’s overall NJ ASK ELA achievement. The model indicates that there is a strong connection between success on the DORA subtests of Comprehension and Vocabulary and a student’s overall scaled score performance on the ELA test of the NJ ASK.