Research Design and Methods

The proposed CMARS is conceived as a web-based, computer administered assessment tool designed to be both student and teacher friendly. It will be student friendly in that each assessment session will feel to the student like he or she is playing a fast paced computer game called, “Right Stuff University.” The student will be engaged in a serried of fast-past, ability tailored subtests requiring no more than 30 minutes to complete in total. After the completion of each subtest, student will be incentivized to continue to achieve peak performance by the sharing of performance data and goal setting. It will be teacher friendly because it will be computer administered, thus requiring little time commitment for the teacher; and yet, will provide immediate, easily interpretable information about student progress. Further, teacher will receive immediate class wide feedback to assist with grouping and instructional target decisions, as well as link to downloadable lesson plans specific to target skills.

Phase I Research Design and Methods

The purposed CMARS assessment will be comprised of five subtests representing the four domains of reading, previous identified. The domain of Word Analysis will be assessed through the spelling subtest. The domain of Fluency will be measured through the connected text fluency and silent reading fluency subtests. The domain of Vocabulary will be measured by the vocabulary subtest, and will include included both general and content area vocabulary. The domain of Comprehension will be measured by the comprehension subtest, and will include several types of comprehension abilities including: determining main idea, making inferences, making critical judgments, and determining cause and effect relationships.

Word analysis subtest. Students will demonstrate if they have fully specified orthographic representation of words in the English language by spelling from among 1,090 carefully selected words that incorporate the various aspects of English orthography. To chose these words, our team first identified approximately five hundred words using grade level word lists for grades 2-14 and analyzed their spellings for number of syllables syllable types, Anglo-Saxon, Greek or Latin roots, affixes, derivatives, inflectional endings, consonant doubling, irregular element, variant spellings, and unaccented syllable schwa. These grade level lists of words were then coded by approximate difficulty with numbers 1 through 5, with one being the most difficult (i.e., having the most elements). Thirty words from each of these difficulty levels was randomly selected from each grade level list resulting in a total of 150 words at grades 4-8. Further, we created an additional 150 items at grade 3, and 75 grade 2 items. An additional 300 items were developed to represent grades 9-14 abilities. Although the difficulty levels for the items were determined based on theory, the Phase II IRT Calibration Study will provide the definitive information regarding the difficulty of each item.

Theory and research. It is known that proficient spellers almost always possess strong word recognition ability, and that good readers typically read at levels near their ability to spell (Foorman & Francis, 1994; Ehri, 2005). Further, better spelling ability is associated with better word recognition, fluency, and comprehension ability (Harn & Seidenberg, 2004). Thus, there appears to be a synergy between spelling and reading (Joshi, Treiman, Carreker, & Moats, 2008; Moats, 2005; Weiser & Mathes, 2009). Learning to spell words and learning to read words are thought to be related like two sides of a coin because they both rely on the same knowledge about the alphabetic system and memory for the spellings of specific words (Bourassa & Treiman, 2001; Ehri, 2000; Ehri & Wilce, 1987; Graham, 2000; Moats, 2000; 2005; Perfetti, 1997). Ehri’s connectionist theory (Ehri, 1997, 1998, 2000) suggests that spelling and reading, although independent skills, develop together reciprocally due to a logical symmetry relationship. Children who spell poorly demonstrate more problems with combining both phonological and orthographic processes together than children who spell well, and children learn about language through print because print provides children with a schema for conceptualizing and analyzing the structure of speech (Ehri, 1998; Ehri 2005). Thus, if one wants to assess how well students are combining phonological and orthographical information with complex multisyllabic words, then assessing student's abilities to spell such words is the logical choice.

Procedure. For this subtest, a line will appear on the screen above the graphic of a keyboard. The computer will ask the student to spell a word. The computer will then say the word in a sentence and repeat the word. Students will use their computer keyboard to type the word. As they type, the letters will light-up on the keyboard that appears on the screen and the letters will appear on the line in the order typed. The purpose of the computer screen monitor is to assist students in keeping their eyes on the screen, rather than looking at their fingers as they type. If a student needs to hear the word again, the student will have the option to push on an icon to have the word repeated. See Figure 3 for an example of the spelling subtest. Words will be selected for the student based on the computer CAT procedure adapting to the child's estimated ability level, regardless of age or grade level. Teachers will be able to access information on the difficulty of the items presented, the types of mistakes their students make, and what type of spelling instruction the child needs for improvement.

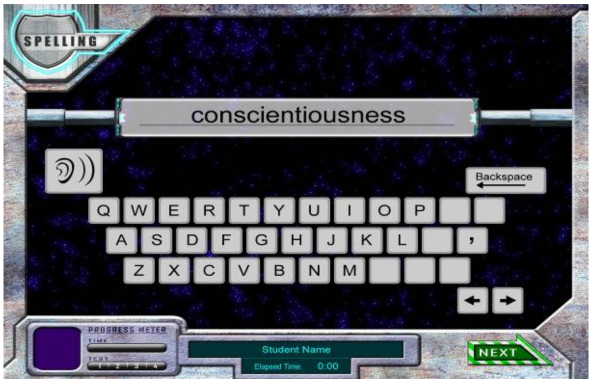
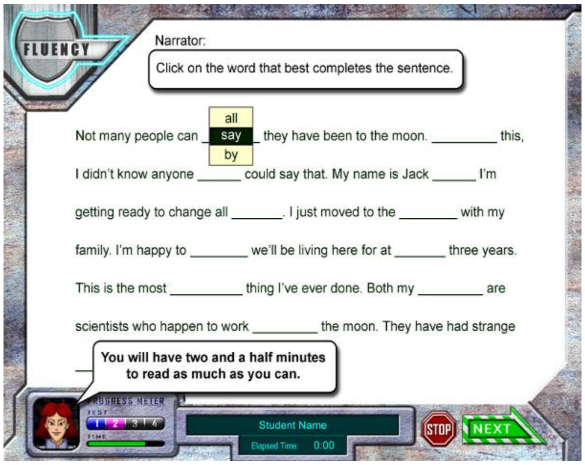


Figure 3. Sample spelling task for CMARS.

Connected text fluency subtest. Students will demonstrate their ability to both read words quickly and monitor for meaning while reading grade level connected text. The subtest will be constructed in a very different manner than the other subtests. Rather than increasing text difficulty across time, students will be assessed on passages of equivalent difficulty to measure growth over time against a constant level of difficulty. We will develop thirty 500 to 700 word stories of near equivalent difficulty for each of the five target grades for a total of 150 stories. Each of these stories will be carefully written to conform to specific word level features, follow linear narrative structure, and have readability according to Flesh-Kincaid and Lexile units for end of grade level the targeted grade. To assess text reading for understanding, a Maze task will be utilized in which every seventh word is left blank from the text. The student will be given three choices for each blank from which to choose the word that works in the sentence. It will be the student's job to read the text, selecting the correct maze response for two and one-half minutes.

Theory and research. Successful fluent readers read connected text with both speed and understanding (Archer, Gleason, & Vachon, 2003; Osborn, Lehr, & Hiebert, 2003). In order to assess the full scope of fluency, measures need to incorporate both speed and meaning aspects of fluency. The Maze task has been shown to be highly correlated to measures of both fluency and comprehension and has high reliability and concurrent validity (Brown-Chidsey, Davis, & Maya, 2003; Fuchs & Fuchs, 1991; Jenkins, Pious, & Jewell 1990; Shinn, Good, Knutson, Tilly, & Collins, 1992; Swain & Allinder, 1996). A similar task was part of the CMERS assessment. Our data confirms that our Maze task, delivered via computer correlates highly to measure of oral reading fluency, comprehension measures, as well as high stakes assessments (Kalinowski, 2009).

Procedure. To complete connected text fluency, the computer will tell students it is time to read a story and review the procedures. The first page then appears, and students perform the Maze task for two and one-half minutes, or until they complete the story. When students complete a page, they click on a button to turn the page and continue. The score obtained from this incorporates the number and accuracy of Maze items completed in the allocated time, as well as accounts for the number of words read between Mazes. This score, which our team formulated for CMERS, has been shown to better correlate to other measures of both DIBELS Oral Read Fluency and comprehension (Lyon & Kalinowski, 2008). An example Maze task authored in the CMARS space exploration theme appears in Figure 4.



Narrator:

Click on the word that best completes the sentence.

all
say
by

Not many people can _ say _ they have been to the moon. _________ this,
I didn't know anyone ______ could say that. My name is Jack ______ I'm
getting ready to change all _______. I just moved to the ________ with my
family. I'm happy to ________ we'll be living here for at _______ three years.
This is the most __________ thing I've ever done. Both my _________ are
scientists who happen to work _________ the moon. They have had strange
__

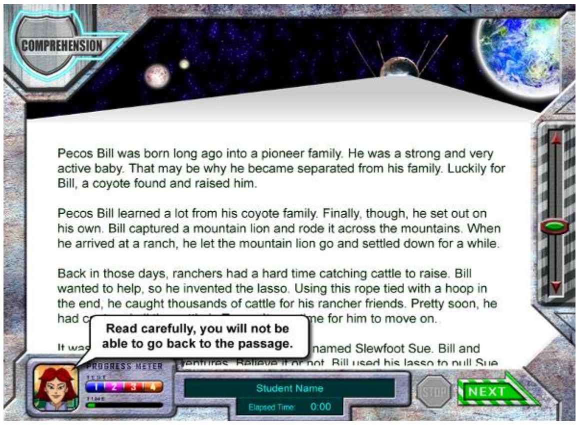
You will have two and a half minutes to read as much as you can.

Figure 4. Sample Maze task for CMARS.

Silent reading fluency subtest. A second aspect of fluency is a student's ability to silently read connected text that is matched to their decoding ability. We propose to measure silent reading fluency by asking students read passages written at a level they can decode while being timed by the computer. Students complete the silent reading of a passage, then press an icon to answer questions about the passage (see the comprehension subtest below). Students will read both narrative and expository passages. While the lower grades will see an equal combination of these two types of text, the upper grades will be reading more expository passages than narrative text. Passages will be composed of varying word counts of 250 to 500 words, with passages written at lower levels being shorter, and more advanced passages longer. There will be 220 total passages created of varying complexity and difficulty, ranging in readability of 2.0 through 12.9 on the Flesh-Kinkaid scale. To assist teachers with assessing the reading ability of their students, we will also Lexile each passage.

Theory and research. Students at about grade 4 transition from gaining more meaning from text read orally to gaining more meaning from text read silently (Prior & Welling, 2001). Not surprisingly, the correlations between traditional Oral Reading Fluency measures and other aspects of reading also become weaker at this time (Brown-Chidsey et al., 2003). Since the ability to read text silently takes on greater importance, and because students in the grades beyond grade 4 are expected to read most of their text silent, the importance of assessing students' ability to read text silently with fluency cannot be overstated. Even so, in our review of the literature, we found that little attention has been paid to this important aspect of reading. Thus, the inclusion of silent reading fluency is, in many ways, experimental. Through the proposed work, we will be able to determine: (a) if silent reading fluency is amenable to measurement in the way that propose, and (b) how well it correlated to the other more established measures of fluency and comprehension. An important aspect of this proposed measure is that we are placing students into text for which they have the ability to actually read the words comprised in the text. Thus, we will be able to ascertain students' silent fluency on text for which they possess the ability to decode.

Procedures. For this subtest, the computer will announce that it is time to read a passage and answer questions (see Figure 5). Students will be told that the computer is timing them as they read the passage, but that they need to read the passage carefully enough to understand the passage without returning to the text. Timing will begin when the passage appears on the page and will end when the student turns the page to begin answering comprehension questions regarding the content of the passage.



Pecos Bill was born long ago into a pioneer family. He was a strong and very active baby. That may be why he became separated from his family. Luckily for Bill, a coyote found and raised him.

Pecos Bill learned a lot from his coyote family. Finally, though, he set out on his own. Bill captured a mountain lion and rode it across the mountains. When he arrived at a ranch, he let the mountain lion go and settled down for a while.

Back in those days, ranchers had a hard time catching cattle to raise. Bill wanted to help, so he invented the lasso. Using this rope tied with a hoop in the end, he caught thousands of cattle for his rancher friends. Pretty soon, he had c me for him to move on.

It was named Slewfoot Sue. Bill and

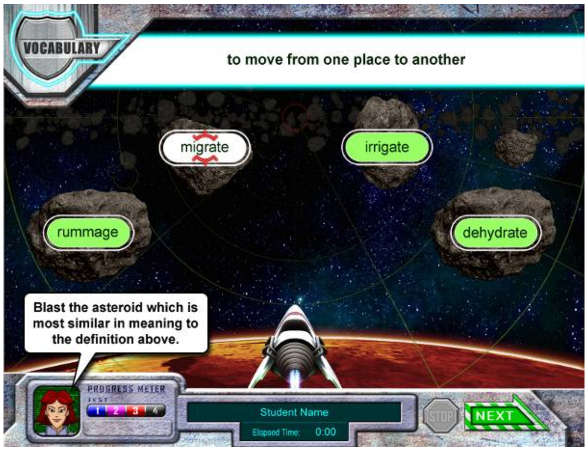
Read carefully, you will not be able to go back to the passage.

Figure 5. Sample silent reading fluency / comprehension passage for CMARS.

Vocabulary subtest. Students will demonstrate their knowledge of word meanings through synonyms or definitions, as well as the ability to infer meaning through context. Four types of questions will be used: (a) select the word that best matches the following definition, (b) select the word that is most similar in meaning to the following word, (c) select the word that best describes the following picture, and (d) select the word that is most similar in meaning to the underlined word. Distracters choices for each word will include words with a similar spelling or pronunciation, antonyms, words an unrelated meaning.

Theory and research. In order to assess students' knowledge of word meaning, we will use decontextualized type of items (synonyms, picture, and definition). However, we also know that students acquire vocabulary best when it is used in a meaningful context. Thus, we also include contextual type of questions, in which students must infer the correct meaning of a word based on its use in a sentence. We have chosen passive recognition tasks for our assessment based on reports that the ability to establish the link between word form and word meaning is the most important component of word knowledge (Laufer et al., 2004; Read, 2007)

Procedures. Throughout the vocabulary assessment, there will be a mix of general vocabulary words and content vocabulary words. The narrator will read the stem for each item. Students can choose to hear the word choices by scrolling over each word on the screen. Students will choose among four possible answers by clicking their mouse on their selected answers. See Figure 6 for an example. The computer CAT program will match the difficulty of the items to the abilities of the students regardless of their age or grade level. Teachers will be able to access reports of their students' progress and needed areas of vocabulary instruction.



to move from one place to another

Blast the asteroid which is most similar in meaning to the definition above.

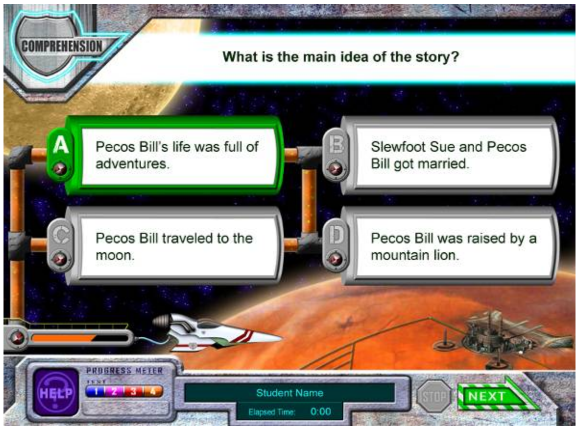
Figure 6. Example vocabulary item for CMARS.

Reading comprehension subtest. The objective of Reading Comprehension subtest will be to determine how well children are processing text of increasing difficulty for meaning. We are constructing 220 graduated passages (ranging in readability of 2.0 through 12.9 on the Flesh-Kinkaid scale) that students will initially read silently. This will also allow us to simultaneously assess silent reading fluency. After reading, students will answer a series of four multiple choice questions. Passages will be a mix of narrative and expository text and will target main idea, cause/effect or problem/outcome, inference, and critical judgment of the text. The underlying theory driving this assessment is that comprehension requires both low level and high level processing of text information. It is in the higher level processing that the deeper message of the text comes forth. Thus, the reading comprehension subtest is being crafted to assess higher cognitive levels of comprehension with the goal of constructing questions that are both conceptually and instructionally valid.

Theory and research. The proposed view of comprehension aligns with our most current understanding of reading comprehension. Higher level processing of text is defined as the reader's ability to determine the overall idea of the passage, differentiate and switch between broader and narrower concepts (essence vs. details), inhibit irrelevant information from intruding upon meaning, monitor comprehension, reason, make inferences, and integrate information into long-term memory (Gamino & Chapman, in press; Kintsch, 1998; Oakhill, Hartt, & Samols, 2005; Sesma et al., 2009; Williams, 2003; Yuill & Oakhill, 1991).

In constructing our items, care is being taken to assess students' coherence of knowledge generation (Kintch, 1998), or the ability to make higher-level links between individual sentences to establish local coherence (i.e., cause/effect and inference question types) and to integrate new information into existing representations to establish global coherence of text (i.e., main idea, problem/outcome, and critical judgment question types) (Cain & Oakhill, 1999; Cain, Oakhill, Barnes, & Bryant, 2001; Oakhill, 1982; Wixson & Peters, 1987). Further, all questions are being designed to be dependent upon information in the passage in order to avoid the testing of background knowledge and having questions that can be answered without reading the text. This situation has been a pitfall of other well-known tests (Keenan & Betjemann, 2006). All answer choices (i.e., correct answer, two distractors, and wrong answer) relate to the passage in some form. Also, because proficient memory has been associated with reading ability and skilled text comprehension (Cain, 2006; Daneman & Merikle, 1996; Sesma et al., 2009; Swanson, Howard, & Saez, 2007), the text will not be available to students when they are answering questions. However, specific details that do not add to an understanding of the general or global coherence of the passage will not be questioned. Thus, once students turn the page to begin answering questions, they cannot see the passage again. Last, we are writing passages that include a range of structures found in both narrative and expository text since comprehension failure has been linked to inadequate knowledge about how texts are structured (Perfetti, 1994). Understanding children's deficiencies in different types of text structures will help when intervening.

Procedures. To complete the comprehension subtest, students will first read a passage that appears on the screen (see Figure 5). The computer will tell them to read the passage for meaning. When they are ready, they will turn the page and the first of 4 questions will appear. When they complete a question, the next question will automatically appear. During the test, students will not be allowed to go back to review the passage. All assessment items will be multiple choice, allowing the student to select from four possible answers. Students will select their answers by clicking their mouse on their selected responses. See Figure 7 for an example of a comprehension item. Teachers will be able to access information of the student's text level, such as overall performance in comprehension based on the student ability index score. Teacher reports will include diagnostic information about skill specific deficits and recommendations for interventions to meet deficiencies.



What is the main idea of the story?

A Pecos Bill's life was full of adventures.

B Slewfoot Sue and Pecos Bill got married.

C Pecos Bill traveled to the moon.

D Pecos Bill was raised by a mountain lion.

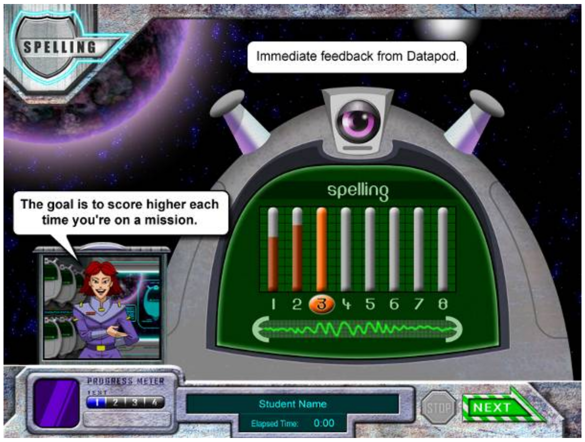
Student Name

Elapsed Time: 0:00

Figure 7. Sample comprehension item for CMARS.

CMARS data management system. From our success commercializing CMERS, istation has the technical experience and savvy to host the backend data management aspects of CMARS. istation already captures data on every key stroke and mouse click a child makes while in CMERS, and the same will be true for CMARS. This "clickstream" data is used internally for system management and diagnostics, as well as externally for reporting purposes.

Student level data. At the end of each session, the computer immediately shows the student his or her graph for each subtest, showing performance on the current subtest (see Figure 8). The graph is set up so that the x-axis shows data points across the year and the y-axis shows relative scores.


Immediate feedback from Datapod.

The goal is to score higher each time you're on a mission.

Figure 8. Sample student feedback for CMARS.

While the graph for each subtest is the only feedback information shown to children, both graphic and skills analysis information is provided to the teacher with an accounting of (a) the skills for which the child has already demonstrated mastery, (b) the skills on which the child is being assessed, (c) performance details on the skills being assessed. Scores reported will include: (a) an IRT-based ability index that represents an estimate of the child's absolute level of ability in a given domain. Because it is not restricted to age or grade levels the ability index can be used to show growth over time, (b) a relative class score representing the quartile range the child is in for her class (i.e. bottom 25%, 25th to 50th percentile, 50th to 75th percentile, and top 25% of the class), and (c) a normed-based percentile ranking comparing the ability scores to a large, nationally representative sample.

Class level data. Because this software is being designed for class wide implementation, the existing CMERS data management and reporting system will be utilized, allowing teachers to examine both individual child and classroom level data. Each individual child's data is recorded in a classroom file. This will allow the computer to aggregate data in the class and generate class level reports. Data for the class will be displayed in rank order form using the most recent data. This is presented in four columns: Mastered, Above the Mean, Below the Mean, and Not Yet Included. Under each column, the names of appropriate children are listed in rank order from highest to lowest. From this list, a teacher can automatically transfer to an individual child's file by clicking on the child's name. For children who are included in the subtest, the child's most recent score appears next to their name. Next to the score appears the child slope or trend score. Children whose slopes are near zero or negative are highlighted to alert the teacher to attend to their academic needs. Likewise, if a child has been designated as requiring "intensive care," the word help appears in red letters next to that child's name. For a child to be designated as requiring intensive care, he or she have must be scoring in the bottom quartile based on our norm reference sample on two consecutive assessments and have slopes are near 0 or negative.

Higher level data. Likewise, a data-gathering feature that will aggregate the data from several classrooms will be developed based on existing code from CMERS. We see this feature as most pertinent to district and building level personnel such as principals, reading specialists, school psychologists, grade level lead teachers, or language arts coordinators. Since CMARS, like CMERS, is web-based, aggregated reports can be generated for virtually any desired aggregation level (nation, state, district, building, or grade) and by any sub-population (gender, ethnicity, SES status, etc.).

Teacher resources. Teacher Resources will be designed to assist the teacher to provide targeted and meaningful differentiated instruction. We will develop downloadable materials including: (a) lesson plans for use during teacher directed, targeted instruction, (b) instructional routines to use during whole class instruction including a peer assisting leaning routines, (c) cooperative learning routines with their associated materials

Item authoring and delivery. CMARS items will be authored using technology and tools previously developed to create CMERS, a web-based reading assessment instrument for the primary grades. Much of the work has already been completed to develop the content of the items and the appropriate look-and-feel for the item types in the four domains of middle school reading. We will rely on the experience gained from our work with CMERS to integrate the items into the existing delivery framework.

istation has developed and successfully implemented state-of-the-art technology for the development, integration, and distribution of educational content through is patent pending Infinity engine which includes: (a) a smart prediction mechanism that downloads the smallest subset of multimedia assets required for each student, (b) the use of peer-to-peer networks to localize request for assets, definitions, and run time engine updates through replication of critical and encrypted student information to multiple computers within a school, resulting in lower bandwidth requirements and system fault tolerance, a feature that ensures service is uninterrupted even during a network failure, and (c) the optimization and reuse of multimedia assets to improve system performance. istation's technology enables high-quality multimedia content to be delivered through the Internet without increasing a district's or school's telecommunications needs or costs. Dedicated servers and infrastructure are *not* required and administration is minimal.

In Phase I, the overarching goal of CMARS is to develop and author items appropriate to measure comprehensive reading ability for all students in grades 4-8. Much work has already been performed towards meeting this goal, including a broad literature review on existing assessments and item types, the design of an engaging and age appropriate theme for the assessment, and the development of many of the 3,100 items and associated reading passages. If awarded this grant, we will focus the Phase I resources on authoring the items into the commercially successful CMERS framework. After completion of Phase I, the items will be ready to be delivered to students in order to establish item-level difficulty and discriminability parameters. After establishing these parameters, the items can then be integrated into a computerized adaptive testing (CAT) environment for production and commercialization.

Phase II Research Design and Methods

The goal of Phase II is to take the items authored in Phase II and integrate them into a computerized adaptive testing (CAT) framework for commercialization (i.e., Phase III). In order to satisfactorily complete the proposed work for Phase II, Phase II will be divided into two studies: (1) the IRT calibration study, and (2) the reliability and validation study.

Prior work with CMERS. In the development and commercialization of CMERS, both an IRT Calibration Study and a Reliability and Validity Study have already been successfully completed using the same partnership proposed in this grant. In 2007, 1,650 CMERS items were delivered to 1,750 kindergarten through grade 3 students from ten school in two North Texas school districts by researchers at Southern Methodist University (SMU). Item response theory (IRT) calibration analyses were performed by researchers at SMU and istation to establish item-level parameters. Subsequently, the items were programmed into a CAT framework for commercialization. In 2008-09, CMERS was delivered in a controlled study to over 400 kindergarten through grade 3 students, along with well-regarded measures of reading ability, to establish reliability and validity evidence (Kalinowski, 2009). Since then, CMERS has been well received by teachers, districts, and state agencies as a respected instrument for continuous progress monitoring of early reading skills. We will use the experiences gained from commercializing CMERS to successfully bring CMARS to the market.

IRT calibration study. An IRT calibration study will be used to determine the item parameter estimates for the pool of items used with CMARS. As with CMERS, a two-parameter logistic (2PL) model will be used to allow for both the item difficulty parameter, as well as the item discrimination parameter to vary by item. Equation 1 illustrates the 2PL model predicting the probability of a correct response to item *j*:

$$P(X_j \mid \theta, a_j, b_j) = \frac{1}{1+e^{-a_j(\theta-b_j)}}, \tag{1}$$

where θ is the person location parameter (i.e., ability), and b_j and a_j are item *j*'s location parameters (i.e. difficulty and discrimination parameters, respectively; Lord, 1980). The 2PL model will be used for the item types having dichotomous responses, such as vocabulary items. For item types with polytomous responses (i.e., correct, partially correct, and incorrect), such as spelling and comprehension testlets, we propose to use Muraki's (1992) generalized partial credit model, also known as the two-parameter partial credit model (2PPC) model (Yen, 1993) as a natural extension to the 2PL model.

Research design. To determine item-level parameters as well as address the model assumptions, a nonequivalent multi-group IRT calibration study has been developed. Students will be recruited from local area schools in much the same way as with the previous CMERS study. However, given the increased number of items, across multiple grade levels (2-14), only a portion of items will be given to students at each grade level. See Table 1 for the proposed distribution of items to students.

Table 1

Nonequivalent, Multi-group Design for IRT Calibration

		Item difficulty (estimated grade level)						
		2-3	4	5	6	7	8	9-14
Students (actual grade level)	8					X	X	X
	7				X	X	X	
	6			X	X	X		
	5		X	X	X			
	4	X	X	X				

Note. An "X" represents a group of students taking items at a particular level of difficulty.

Given field tests of the items types for the four domains, we estimate that students will need sixteen sessions of between 35 and 40 minutes to complete all of their allotted items from all subtests. As per our previous IRT study for CMERS, we will ensure that adequate time will be allocated for the collection of data, including extra sessions to address absentees and student functions interfering with the study. Trained graduate research assistants will supervise the students as they respond to the items, cautiously guarding against off-task behavior that might impact authentic responses to the items.

Our goal is to recruit approximately 400 students at each grade level (4-8), for a total of 2,000 students in the study. Given this design, each item will have between 400 and 1,200 responses, which is adequate for accurate parameter estimation in a 2PL IRT model (de Ayala, 2009). To estimate the item parameters, BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) will be used for dichotomous item types, and MULTILOG (Thissen, Chen, & Bock, 2003) will be used for polytomous item types. Both IRT programs use marginal maximum likelihood estimation (MMLE) to maximize the person response vector across both the item difficulty and discriminability dimensions. For example, Equation 2 represents the probability of a response vector of dichotomous items, **X**, in an instrument of length L,

$$P(\mathbf{X} \mid \theta, \mathbf{J}) = \prod_{j=1}^{L} p_j^{\,x_j} (1 - p_j)^{1-x_j} \,, \qquad (2)$$

where the probability of a set of responses is conditioned on the person's ability (θ) and the matrix of item parameters, **J** (i.e., the collection of a_js and b_js for each item, *j*). In MMLE, an unconditional, or marginalized, probability of a randomly selected person from the population with a continuous latent distribution is specified as an integral function over the population distribution (Bock & Aitken, 1981). Subsequently, the resulting marginal likelihood function undergoes maximum likelihood estimation (MLE) by BILOG-MG (and MULTILOG) to generate item parameters.

Model assumptions and analyses. The 2PL model is predicated on a unidimensional latent space and conditional independence assumptions (Lord, 1980). Model-data fit statistics generated from BILOG-MG and MULTILOG will be used ascertain the dimensionality of the data. Yen's (1984, 1993; Kim, de Ayala, Ferdous, & Nering, 2007) Q_3 index will be used to evaluate conditional item dependence. Additionally, empirical item characteristic curves (ICC) will be compared against predicted plots via reported χ^2 statistics to identify items that do not have desirable psychometric properties.

Further, each item will be analyzed for bias across subpopulations. First, functions within BILOG-MG and MULTILOG will be used to detect group differences with regards to gender and ethnicity. Items identified as exhibiting differential item functioning (DIF) will be reviewed by a panel of reading experts to determine whether the source of the item's differential performance is relevant to the construct being measured. If an item is determined to have logical evidence of bias, it will be removed from the pool of items.

Integration into CAT. After the items have been calibrated and biased items eliminated from the pool, programmers at istation will integrate the items into a computerized adaptive testing (CAT) framework based on the commercialized product, CMERS. As with CMERS, CMARS CAT will use the Bayesian expected a posteriori (EAP) strategy to estimate a person's ability location (Bock & Mislevy, 1982). Because EAP can be used to obtain location estimates for all response patterns, including zero and perfect response vectors, an estimated ability will be produced after every response. Subsequent items presented to the students will be selected based on the maximum information produced when the difficulty of the item is nearest to the ability of the test taker. As with CMERS, the CAT stopping criterion will be a minimization of the standard error of the ability estimate. After integration and quality assurance testing, the CMARS product will effectively be production-ready. However, before commercialization occurs, CMARS will undergo a study to assess the reliability and validity of the data.

Reliability and validity study. A reliability and validity study will be used to determine the consistency and accuracy of CMARS data as compared to other widely used measures of reading ability.

Research design. Approximately 500 students from grades 4-8 will be recruited from multiple North Texas school districts to participate in the study. Students will be escorted to the school's computer lab and administered all assessments by trained graduate research assistants every two to three weeks. Depending on the time of year the study takes place, it is expected that this study will take between two and four months to complete (i.e., holiday breaks may extend the length of the study).

Reliability. Cronbach's (1951) coefficient alpha is often used as an indicator of reliability across test items within a testing instance. However, alpha assumes all students in the testing instance respond to a common set of items. Due to its very nature, students taking a CAT-based assessment, such as CMARS, will receive a custom set of items based on their initial estimates of ability and response patterns. The IRT analogue to classical internal consistency is marginal reliability (Bock & Mislevy, 1982). In essence, marginal reliability is a method of combining the variability in estimating abilities at different points on the ability scale into a single index. Like Cronbach's alpha, marginal reliability is a unitless measure bounded by 0 and 1, and it can be used with Cronbach's alpha to directly compare the internal consistencies of classical test data to IRT-based test data.

To establish test-retest reliability evidence, Pearson product moment correlation coefficients between multiple CMARS testing instances will be computed.

Validity. In addition to taking CMARS every two to three weeks, students will be administered one external measure. To reduce ordering effect, a Latin squares design will be utilized. Although the list of external measures has yet to be finalized, a preliminary list includes Woodcock-Johnson Tests of Achievement (WJ-III ACH; Woodcock, McGrew, & Mather, 2001), Woodcock Language Proficiency Battery-Revised (WLPB-R; Woodcock, 1991), Wechsler Individual Achievement Test (WIAT-II; Wechsler, 2005), Iowa Tests of Basic Skills (ITBS; Hoover, Dunbar, & Frisbie, 2007), and Stanford Achievement Test Series (SAT10; Pearson Assessments, 2009). Pearson product moment correlation coefficients between CMARS data and external measures will be computed.

Given that the study will take place in Texas schools, it is proposed that the study include the release of the high-stakes, end-of-year Texas Assessment of Knowledge and Skills (TAKS; Texas Education Agency, 2003) reading scores for the study participants. Predictive validity evidence will be computed, including receiver operating characteristic (ROC) analysis, to determine how well CMARS predicts success on TAKS. Although the results would not be generalizable to other state tests, (a) a high proportion of the existing CMERS market is Texas school districts, so the information would be valuable to CMARS commercialization efforts, and (b) the results would yield information as to how well CMARS might predict end-of-year tests for other states given the similarity in learning objectives between state reading assessments.

Reporting. During the reliability and validity study, teachers, reading coaches, and principals will be selected and asked to use the CMARS reporting interface. As they navigate the reports, they will be asked two driving questions: (a) What do you see? and (b) What would you do with the information presented? This qualitative look into our reports will help us understand how the student-level and aggregate reports are used by the stakeholders. If trends in the data can found regarding areas of the reports that need to be addressed, changes will be proposed to engineering prior to commercialization.

In Phase II, the main goal will be to determine the psychometric properties of the pool of items developed and authored in Phase I, program them into a CAT framework, and collect reliability and validity evidence for the resulting CMARS instrument. Further, focus groups on select stakeholders will help istation determine if the existing reporting interface is satisfactory for making continuous progress monitoring decisions for their students. After completion of Phase II, CMARS will be ready for commercialization into established markets that have embraced its predecessor, CMERS.