Validity and Reliability Testing of the Iowa Assessment of Skills and Knowledge for Automatic Word Recognition

Keith S. Apfelbaum
Carolyn Brown
Jerry Zimmermann
Foundations in Learning, Inc.

Technical Report Prepared for Foundations in Learning, Iowa City, IA
March, 2018

For more information, contact:
Keith S. Apfelbaum
Director of Research, Foundations in Learning, Inc.
2261 Crosspark Road, Suite 123
Coralville, IA 52241
319-333-7788
kapfelbaum@foundations-learning.com

Introduction

Reading is a critical skill for academic and life success. However, a startling number of students in the United States struggle to read throughout their formal schooling and in the work place. The 2013 National Assessment of Educational Progress (NAEP) found that around two in three middle school students were below grade-level proficiency in reading (US Dept. of Education, 2013). More troubling still, additional research suggests that around half of these students have deficits in foundational reading skills, such as decoding and fluency, which they should have mastered during elementary school (Cirino et al., 2013; Hock et al., 2009). This finding highlights a daunting challenge for many students; they never gained the requisite knowledge to effectively read and are no longer in an instructional environment that is organized to teach these skills.

A lack of foundational reading skills in middle school portends major difficulties moving forward for several reasons. Foremost is that middle school students must read for content and acquire new knowledge to academically advance. However, those students who lack the basic reading skills to read with fluency must devote their cognitive resources to decoding rather than comprehension. Further compounding this difficulty, middle school teachers lack training in identifying basic reading difficulties and in teaching these foundational skills. Finally, middle school students have extremely limited time to recover from their deficits in reading skills. There is little time within a school day in middle school to devote to basic reading intervention, and the window to develop and successfully use these skills before they enter high school is rapidly closing. Efficient identification and effective intervention are critically important to the academic success of these students who struggle to read.

Summary of the problem: Current assessments typically identify whether basic word-level problems are present, but do not differentiate between what students *know* and whether they can efficiently *use* that knowledge. For example, some struggling students can decode words, but they cannot automatically use and generalize these skills. Others can neither decode nor generalize word recognition skills. Without more differentiated diagnostics followed by efficient and effective interventions for these gateway reading skills, these struggling middle school readers will never attain grade level fluency and comprehension.

A new approach to evaluate both knowledge and use of skills: There is a clear need to develop a diagnostic that can pinpoint the gaps in individual students' profiles of word-level knowledge (decoding) as well as their use of this knowledge (generalization and automaticity) to best inform intervention instruction. Such an approach to assessment demands an understanding of the componential nature of basic reading and the valid and reliable measurement of these parts.

Grapheme-phoneme correspondence (GPC) regularities of the language (i.e., how specific letters link to sounds) include vowel and consonant classes, as well as subclasses within these (e.g. vowel digraphs), and specific correspondences within the subclasses. Evaluation of GPC knowledge has been primary to the identification of decoding problems, which often triggers phonics intervention for students with deficits.

However, GPC knowledge alone is insufficient for effective, fluent reading. A fluent reader must also be able to deploy this knowledge rapidly and effortlessly to quickly read and comprehend. A student who must slowly sound out every word will struggle to read connected text, just as a student who does addition by counting on her fingers will struggle to complete more complex math problems. Skillful deployment of GPC knowledge is thus critical to effective reading. Skillful use of phonics is also not a unitary construct. Instead, there are several ways that this deployment should be optimized for a skilled reader. For instance, an effective reader should be able to generalize their knowledge to new contexts, including to nonsense words and to words of different length. An unskilled reader may be able to memorize the spelling of common words, but they are likely to struggle when they need to decode less common words or words they have never encountered before (such as nonsense words). Additionally, a skilled reader needs to deploy her knowledge quickly and effortlessly; reading fluently for content demands the ability to decode many words in rapid succession, and cognitive resources are best devoted to comprehension, rather than the procedural aspects of decoding.

This latter skill (rapid deployment of reading knowledge), although a typical goal of reading training and intervention, is rarely directly investigated at a mechanistic level. Extant assessments gauge constructs like fluency, or students' ability to read connected text in a smooth, rapid fashion. However, such assessments lack specificity of what underlies this ability. A componential approach, such as that taken in the current work, asks what enables the reader to effectively deploy her knowledge rapidly and effortlessly. The construct of *automaticity* is relevant to reading, just as it is to skill development in many domains. Automaticity refers to the ability to complete a procedural skill with minimal conscious effort. This construct has been investigated extensively in various cognitive science domains, particularly in motor skill learning (Wulf, Shea, & Lewthwaite, 2010). Numerous studies show dissociations between explicit learning of a skill and robust encoding of that skill for automatic use. Critically, these two goals often benefit from different forms of training or instruction, suggesting that automaticity is not a given outcome of acquisition of knowledge. Reading may show the same dissociation. Separate assessments to measure skill automaticity and knowledge of GPC classes would provide a more comprehensive measure of a student's reading abilities, and thus a more appropriate roadmap for effective intervention that acknowledges the distinction between knowledge and skilled use of this knowledge.

The Iowa Assessment of Skills and Knowledge for Automatic Word Recognition and Decoding (iASK)

The assessment detailed in this document was designed to precisely identify gaps in decoding skills as well as deficiencies in automaticity of word recognition for middle school students. It provides a level of specificity in the differentiation of foundational reading needs that allows teachers and interventionists to attack these reading roadblocks with more precisely targeted intervention. This assessment embraces the componential nature of foundational reading skills, and assesses both knowledge (decoding) and skills (automaticity). Further, within these components, individual subcomponents are measured to determine how well students perform with different GPC classes, in different contexts and on different types of tasks. The output of the assessment provides sufficient specificity of information to inform intervention at the individual level.

An iterative process was undertaken to optimize the design of the assessment so that only highly informative tasks and items were included. The final form of the assessment was validated as an effective measure of middle school student reading performance, both in terms of knowledge and skills, and it was shown to be a highly reliable tool for assessing student profiles of performance. This technical report details this validation process for the final form of the assessment, as well as a short-form screener developed to predict broader student deficits in a short time.

iASK consists of a series of tasks that assess student reading abilities in a variety of ways, using a wide range of different items. This diversity allows measurement of numerous subcomponents of reading skills. Using a computer-based platform of multimodal tasks, iASK is designed to maintain student engagement, minimize the need for teacher training, and ensure fidelity of implementation across students. The completed version of the full diagnostic assessment consists of 708 total trials distributed across multiple forms of five primary tasks. Students complete this assessment in approximately 90 minutes, typically completed over three to four sessions. A unique username and password for each student automatically tracks performance across sessions. Additionally, a short-form screener version of the assessment is available, which provides a coarser approximation of student performance on overall decoding and automaticity (described in detail at the end of this report). The screener can quickly distinguish whether a struggling reader is likely to have foundational deficits sufficient to engage in the complete diagnostic. This screener consists of 144 trials and can be completed in a single session of approximately 15-20 minutes.

Validating iASK

We followed an iterative process to validate iASK as an effective tool for assessing middle school students' reading knowledge and skills. In initial testing, we included an extremely wide range of tasks and items to determine what best predicted student performance. We then honed in on the most diagnostic tasks and items to predict student performance comprehensively, reliably and efficiently. This consisted of two distinct phases of data collection and analysis. The first wave assessed broad feasibility and usability of the overall design and particular tasks used in iASK. The second phase included multiple consecutive waves of data collection to optimize the design of the assessment. Through these waves, we winnowed the set of tasks and items to the fewest necessary to maintain strong predictive validity and reliability of the critical constructs of interest. With each iteration, we investigated multiple facets of the assessment, including: whether each task was easily understood by the students; whether each task was predictive of relevant outcome variables; whether specific items within tasks were more predictive than others; and which specific factors (e.g. which GPC regularities; words vs. non-words) were most predictive within each task.

Throughout these successive waves of data collection, we analyses allowing for these advancements relied on a series of regression and IRT models to determine how individual factors predicted various outcome measures, with separate analyses investigating how removal of tasks and items would impact these measures. These analyses identified several tasks that were either uninformative about foundational reading skills in this group, or were extremely highly collinear with other tasks, and thus unnecessary for diagnostic purposes. We also identified the most relevant factors to use for counterbalancing to ensure meaningful and valid output, and optimized the overall design for ease of student use. This technical report focuses on the validity and reliability testing conducted on the final form of the assessment developed from the outcome of these iterative measures.

Structure of the Assessment. The final tested form of the iASK assessment included five base tasks: Change the Word (CtW): students are shown a word, and then asked to change it by adding a new letter(s) (e.g. "change *bat* to make *bet*"). Students then have to choose the appropriate letter from eight alternatives to complete the change. Fill in the Blank (FitB): students are shown a word that is missing a letter(s). They hear the word spoken, and must choose the missing letter(s) from eight alternatives. Picture Matching (Pic): the student sees a word printed on the screen, and must choose the picture that matches this word from four alternatives. Foil pictures are chosen that have overlap in their spelling (e.g. foils for *face* are *fish*, *piece* and *flake*). Rhyme Identification (Rhyme): the student sees a word printed on the screen, and must choose a word that rhymes with it from a set of six alternatives. Foils have overlap in spelling. Word Verification (Verify): the student sees a word printed on the screen, and hears one spoken. The student must signify whether the printed word is the same as the spoken word. In half of trials, the spoken and printed word mismatch; the mismatching printed word is chosen such that it has close orthographic overlap. Several variants of these tasks were included in the assessment; for example, FitB, Pic and Verify all included both masked and unmasked versions of the task. In masked versions of tasks, the printed words were covered up after 90 msec of viewing time, to force students to rely on automatic processing to read them, whereas unmasked variants had no time pressures. Additionally tasks could have either monosyllabic or multisyllabic items; for multisyllabic blocks, words were balanced between one, two and three syllable items. A full list of the tasks and their variants included in this form of the assessment is presented in Table 1.

Table 1: Tasks and task variants used in iASK, along with trial counts for each variant.

Task Name	**Blocks**	**Trials per block**
Fill in the Blank Monosyllabic	4	18
Fill in the Blank Multisyllabic	6	18
Change the word	4	18
Verify Monosyllabic	3	16
Verify Masked Monosyllabic	3	16
Verify Multisyllabic	3	16
Verify Masked Multisyllabic	3	16
Rhyme Identification	3	16
Rhyme Identification Masked	3	16
Find the Picture Monosyllabic	2	18
Find the Picture Masked Monosyllabic	2	18
Find the Picture Multisyllabic	3	16
Find the Picture Masked Multisyllabic	3	16

The total number of trials per block varied to allow finer control of experimental design and decrease the total duration of the assessment. For example, rather than including separate blocks to assess consonant and vowel knowledge, trials assessing these were interspersed within the same block. This allowed greater design economy, as fewer total blocks of trials were necessary (eliminating delays that occur between blocks, as children select the next block). Additionally, this allowed more straightforward counterbalancing of lexical factors; we could closely control the amount of exposure to different vowel and consonant types in tandem by balancing these within a block.

Blocks of tasks were distributed across “levels,” which were instantiated as six different destinations on an island map (see the Teacher’s Guide for depictions of map destinations, the task selection screen and other aspects of the interface). Each destination included seven blocks of trials represented as a cloud hovering over the destination; the student had to complete all seven blocks within a destination before moving on to a new destination. Students progressed through the blocks at their own pace; they were free to choose which destination to complete, and then within that destination, they could select the order of tasks (embedded in “clouds”). After completing all blocks within a destination, students returned to the main level selection

Table 2: List of GPC classes included in iASK, with examples of items from each class.

GPC Class	Description	Examples
CS1	Simple consonants at the start of a syllable. These are single-letter consonants that produce a single sound.	**C**at; **m**oon; buil**d**ing
CDG1	Consonant digraphs at the start of a syllable. These are multi-letter consonants that make a single sound (e.g. SH, CH, WR).	**Sh**epherd; **kn**ife; **wh**ale
CCL1	Consonant clusters at the start of a syllable. These are multi-letter consonants that make multiple sounds (e.g. CL, BR, SPR).	**Cl**ub; **br**oom; **spl**itting
CS2	Simple consonants at the end of a syllable. These are single-letter consonants that produce a single sound.	Boo**t**; cra**b**; ra**n**dom
CDG2	Consonant digraphs at the end of a syllable. These are multi-letter consonants that make a single sound (e.g. SH, CK, LL).	Doorbe**ll;** du**ck**; cra**sh**
CCL2	Consonant clusters at the end of a syllable. These are multi-letter consonants that make multiple sounds (e.g. SK, LD, TCH).	Ta**sk**; pa**tch**work; gi**ft**
SH	Short vowels. Single-letter vowels that produce their dominant pronunciation.	C**a**t; f**e**tch; p**u**ddle
LO	Long vowels. Vowels with the long-vowel pronunciation driven by a syllable-final silent E.	B**i**t**e**; par**a**d**e**; b**o**n**e**
DG	Digraph vowels. Two-letter vowel combinations that produce their dominant pronunciation (e.g. EA, OA, AI).	B**ai**t; cl**ea**ner; r**oo**t
DR	Diphthong and r-colored vowels. Letter combinations that create either diphthongs or r-colored vowel sounds.	B**ar**k; b**ou**nd; b**oi**ler
DS	Digraph secondaries. Multi-letter vowel combinations that create either secondary pronunciations or exceptions. For example, words when EA makes a short-vowel sound.	H**ea**d; b**oo**k; f**ou**r
SS	Short secondaries. Single-letter vowels that create either secondary pronunciations or exceptions.	B**a**ll; bl**i**nd; f**u**ll

screen to choose their next destination. In a given session, students spent 30-40 minutes working on the tasks, completing as many blocks as possible within that time window; when their time elapsed, they completed the block of trials they were currently in, and then were logged out. Student progress was automatically saved at the end of the session, and they began where they left off the following testing day. In total, students completed 708 trials; this was a substantial decrease in trials from the previous wave of testing, and a more practical length of assessment for classroom time constraints.

The tasks targeted different aspects of reading skills, so item characteristics were balanced separately within each task. For example, generalization to words of different lengths was specifically targeted in FitB, Verify and Pic tasks, so these tasks included a separate counterbalancing for syllable number. The Pic task does not accommodate non-word trials, so this task only used word items. Within each task, counterbalancing was done to optimize balance between GPC classes, word types, word complexity and targeted GPC class, as needed. Adding flexibility in the number of trials per block enabled more precise control of counterbalancing, ensuring sufficient power for each measure while limiting the number of trials needed.

In general, each task was counterbalanced for the targeted GPC classes (Table 2), with these evenly balanced between and within each task by word vs. non-word status (when applicable), as well as their non-targeted orthographic forms (e.g. for trials targeting vowels, the complexity of the consonants in those items was controlled across the vowel GPC classes). Items were also balanced by these factors across the masked and unmasked versions of tasks, as well as across mono- and multi-syllabic versions of the tasks. These counterbalancing approaches ensured that differences between GPC classes or between different forms of the tasks were most likely driven by the assessment-relevant factors, and not by other gross differences between items. In some cases, perfect counterbalancing was impossible, due to limitations of age-appropriate items that met the required criteria (e.g. an item that is readily picturable with specific vowel and consonant GPCs that a middle school student is expected to know). In these cases, items were chosen that were as close to the target characteristics as possible.

Task-specific factors. Specific facets of the tasks were chosen based on the testing conducted with earlier forms of the assessment. This testing led to design choices that proved highly predictive of outcome variables. Specifically, the following aspects of tasks were chosen based on these previous tests:

1) The masking interval. During the masked (or speeded) versions of tasks, students see the printed word for a short interval, after which the word is covered up by a series of # symbols. The masking interval used in this final version of iASK is 90 msec. This interval allows students to view the item, but prevents slow, effortful processing. Our testing showed that this interval decreases student performance, but is not prohibitively difficult. Additionally, this masking interval led to strong predictive validity of outcome measures, particularly those for fluency (more details on this in *Results*).
2) The inclusion of both word and non-word items. The inclusion of non-word items in reading assessment and intervention programs is sometimes controversial. However, we found that inclusion of non-words was highly effective for prediction of student abilities. These items preclude any possible memorization strategies that students could use, and instead force a more pure decoding approach. When analyzing student performance using only the real-word items, we were able to account for significantly less variance in student decoding abilities, as measured by standardized measures of decoding. Having both types of items thus enables a better measure of student abilities.
3) The inclusion of multi-syllabic items. Previous research demonstrates that struggling readers (e.g. those with dyslexia) often struggle with longer words (Hudson & Bergman, 1985; Zoccolotti et al., 2005). However, the evidence that such effects hold for a sub-clinical sample is more mixed. Our initial testing showed that performance decrements

observed in conditions with multi-syllabic items were predictive of outcomes. As such, we included both mono- and multi-syllabic versions of our tasks. Multi-syllabic items were included in separate blocks, which used one-, two- and three-syllable items. These items were separately counterbalanced for lexical factors from the balance used in the mono-syllabic blocks, as some factors were less straightforward to control for (e.g. these items have multiple vowels, and several consonant positions, so it is difficult to assign a single vowel to them). For this reason, analyses of syllable length effects compare within the multi-syllabic blocks, and not directly to the mono-syllabic blocks.

Item selection. Items were selected for inclusion in the assessment based on the difficulty estimates from previous waves of testing. These measures showed which items were at an appropriate level of difficulty to discriminate student abilities (i.e. the items were not so easy that all students got them right, nor so difficult that all got them wrong). We pseudo-randomly assigned words so that predicted difficulty was approximately equal across tasks. To avoid possible item idiosyncrasies, in which specific item selections may under- or over-estimate student abilities because of unintended item-level differences, we generated six full item sets, choosing subsets of items for each that had relatively low overlap between sets. Initial item selection for each iteration was accomplished using a computer script to choose items that met the counterbalancing criteria and were closely controlled for item difficulty. Each of these iterations used a different starting random seed to ensure variability between curricula. Then, we manually checked each curriculum for any issues (e.g. imbalances in task variables that were not optimized in the script; excessive overlap between items in a task). When such issues were encountered, the items were manually replaced for that iteration.

This created six unique curricula, which we could use to determine reliability of our measures independent of specific item choices[1]. Given our constraints to choose items from a confined range of difficulty, needs to balance GPC class and other lexical characteristics, and task-specific constraints (e.g. Pic requires items that have pictures), there was some necessary overlap between curricula. Each curriculum shared a mean of 432.8 items (*SD*=9.39 words) out of the full 708 items with each other curriculum. However, only 131.6 words were shared within the same task (*SD*=11.16). This frequent overlap but in different tasks proved beneficial for analysis; having multiple measures of an item used in different tasks allowed us to assess specific item difficulty more comprehensively while maintaining predominantly independent curricula.

Each of the six curricula had four possible implementations. For the Verify tasks, half of the items had matching audio and printed word, and the other half mismatched. Across implementations, we counterbalanced which items were matching and which were mismatching. Additionally, we counterbalanced the choice of items for masked and unmasked tasks. Thus, we had 24 total possible curricula: 6 base curricula by 2 counterbalances for Verify tasks by 2 counterbalances for Masking tasks.

For the initial round of data collection, students were randomly assigned to one of the 24 curricula, with care taken to balance: gender; grade; school; and standardized test score on the Iowa state reading assessment. For students that returned for reliability testing, their second curriculum always used a different base curriculum from their first. For each initial curriculum choice, the reliability choice was evenly distributed across all other curricula. Additionally, the

[1] These multiple curricula also form the basis of the multiple forms of the final assessment, allowing a single student to complete the assessment at multiple time points and compare performance across these times.

choice of reliability curriculum was balanced by student, grade, gender and Iowa assessment score.

Procedure. Participants were identified based on their scores on the Iowa state standardized reading assessment. Any student who fell between the 10th and 60th percentile was identified as a potential participant for the study, unless they had a clinical reading diagnosis. Qualifying students were invited to participate, and informed consent for participation was requested from their parents. Those that consented to participate took part in both standardized assessment measurements and the full iASK diagnostic. A subset of participants also completed a second wave of iASK using a different curriculum approximately two weeks after their initial testing.

Prior to beginning the iASK assessment, the students completed a battery of standardized assessments of their reading and language skills. These assessments were run by trained research assistants, and they were conducted in the schools with individual children or small subsets (depending on the assessment). Based on analyses of the previous waves of data collection, we chose assessments that best exemplified skills we wanted to measure. To this end, we included the Woodcock-Johnson Word Attack and Word ID subsections as measures of general decoding ability. These were averaged to offer an overall decoding score against which to validate iASK. To assess reading fluency, we included the Texas Middle School Fluency Assessment (TMSFA). We also assessed reading vocabulary and reading comprehension with subtests of the Gates-MacGinitie Reading Tests, and oral vocabulary with the Peabody Picture Vocabulary Tests. This suite of tests offered a fairly comprehensive view of student language and reading abilities, allowing us to assess the discriminant validity of iASK in predicting both decoding and automaticity abilities independently.

Participants. After gathering informed consent, 224 students completed the standardized assessments and the iASK assessment, and thus were included in analyses. These students were all in grades 6-8 at middle schools in the Cedar Rapids (IA) Community School District, and all fell between the 10th and 60th percentile on their yearly Iowa standardized assessment of reading ability. We thus assessed students on the low end of the reading spectrum, but also those in the typical range. This sample allowed us to assess both students who demonstrate reading deficits, as well as those without deficits, to validate that our assessment could predict which students have proficient reading skills. Among the final sample, 131 students also took part in reliability testing, completing a second full run of iASK two weeks after this first testing.

Results. Data were analyzed separately for the separate components of interest: decoding; automaticity (performance in masked vs. unmasked trials); generalization across number of syllables; and generalization from words to non-words. For each of these factors, we conducted a separate set of models, using a subset of tasks and trials that were deemed of primary interest for the construct in question. Details of each model are described below. Within each construct, we conducted several successive steps to achieve interpretable output that is both highly robust and commercially feasible to compute.

First, for any construct that included an unclear number of relevant factors, we conducted hierarchical linear regressions to determine which factors were necessary to predict our outcome measures. For these regressions, we input the different factors into the regression as predictors at different steps, and determined whether their addition explained variance beyond the other factors. We conducted these analyses using both the Woodcock-Johnson (averaged between the two subparts) and the TMSFA as dependent variables (DVs); if the factor predicted significant relevant variance in either, we included it in later analyses. For example, when considering the construct for generalization across words of different syllable length, we had available data from both masked and unmasked versions of the multisyllabic tasks. The output of the hierarchical regressions showed that including the masked versions did not account for additional variance over using the unmasked versions alone. As such, we opted to only include the unmasked versions in later analyses of the word length effect. This approach allowed us to optimize our analyses for each construct to include only those factors that are most informative for that construct. This also allowed us to achieve good discriminant validity between our constructs, by only including the subset of factors needed for each individual construct, rather than incorporating all factors at once.

Once the relevant factors for each construct were identified, we conducted item response theory (IRT) analyses using these factors to determine the appropriate weights for each factor. We separately computed IRT estimates for the initial data and for the reliability data. Each analysis included trial accuracy as the DV. Fixed effects were selected for all variables that were not of interest, while random effects were used for both targeted variables of interest and the variables not of interest. The option to not include fixed effects for variables of interest ensured that our participant-specific random effects accounted for all relevant variance in the variable of interest, without the fixed effect biasing these random effects. That is, by basing participant predictions entirely on their random effects for the variable of interest, we achieved a much more precise estimate of that individual's performance, without allowing group means to bias this estimate. Simultaneously, including fixed effects of non-targeted variables removed spurious variance from other factors in the assessment. For example, some tasks were overall more difficult than others; adding fixed effects for tasks ensured that task difficulty was accounted for in the model before estimating participant effects for variables of interest. Along these same lines, for constructs in which different items contribute to different cells (such as the word vs. non-word model), we did not include random effects for items, as these random effects would subsume overall variance that is better accounted for by the participant effects of word vs. non-word.

The output of these IRT models allows us to estimate participant performance on the different constructs, while accounting for external variability from factors like task differences or surrounding orthographic context (e.g. when estimating vowel GPC decoding abilities, we could account for the consonant context). As such, these estimates are highly effective measures of student performance. Additionally, the measures were generally highly predictive of the outcome measures. These findings validate our analytic approach, and iASK's measures more generally, by demonstrating that they quite effectively predict performance on formative assessments of student reading ability. More importantly, they do so while producing individual profiles of performance for particular aspects of these abilities.

However, IRT analyses are computationally expensive and slow, and thus are not feasible for use on a commercial scale; to provide timely reporting of student performance, it is necessary to find a more computationally tractable approach to estimating student performance. To this end, we conducted multivariate multiple regression analyses to determine how to estimate the IRT output given the measures of student performance. These analyses identify which factors are significantly related to the IRT estimates, and provide weights for each factor. On a practical level, these analyses allow us to compute a simple linear equation for each of our DVs that closely approximates the IRT outputs; these equations are computationally simple, and so are more appropriate for commercial utilization.

For the word/non-word models, we considered the effects of word vs. non-word in each of the other constructs; because words and non-words do not constitute a clear distinction of relevant tasks or task characteristics, there were a number of possible model structures that could be used to estimate word/non-word performance. To ensure we used the most effective estimate, we compared the output of the word vs. non-word distinction in each of the decoding, generalization across syllables, and automaticity models. For each model, we constructed a sub-model to target the word/non-word effect. We then examined:

1) How well the estimates of word/non-word predicted the standardized measures;
2) How reliable the word/non-word estimates were.

We found that the word/non-word effect from the decoding model was fairly unreliable (r=.16) and showed only weak correlations with the outcome measures (with TMSFA: r=.38; with WRMT: r=.22). Meanwhile, both the generalization and the automaticity models produced more reliable (r=.37, r=.52, respectively) and more predictive (with TMSFA: r=.54, r=.55, respectively; with WRMT: r=.58, r=.45, respectively) estimates of performance. Given the tradeoffs between the two models, we next considered a mean of estimates from the two models. This estimate proved more reliable (r=.62) and more predictive (TMSFA: r=.61; with WRMT: r=.55). Thus for our model of word/non-word performance, we used the mean estimates from these two models as our estimate of performance.

We conducted a separate multivariate multiple regression for each of the following constructs: decoding--vowels; decoding--consonants; generalization--word vs. non-word; generalization--number of syllables; and automaticity--masked vs. unmasked. For each of these models, we included only data from the initial assessment. Multivariate regression predicts multiple DVs simultaneously, and thus accounts for shared variance between the DVs. For DVs, we used the estimated outputs for each construct of interest from the IRT models. For those students who also completed reliability testing, we gauged the reliability of these estimates without having their reliability data contribute to the coefficients.

Since a typical multivariate regression computes coefficients for each of a given set of factors, we needed to determine the necessary set of factors as well. Thus, we utilized an iterative process to conduct a series of multivariate regressions, and determine which factors were necessary. We began with a set of factors that we deemed essential; these factors were directly relevant to the DVs being estimated. We then ran successive regressions adding additional factors and determining whether overall model fit was significantly improved. Because this process included multiple comparisons, and because of the fairly large sample size included in the models, we opted for a conservative significance criterion for including additional variables ($p<1*10^{-10}$). Additional analyses with more stringent significance thresholds produced extremely comparable results.

These analyses yielded outputs that predicted the output of the IRT analyses extremely closely. The correlations between each DV within each model and the corresponding IRT estimate are presented in Table 3. Overall, these strong correlations demonstrate that our multivariate multiple regression approach was effective in estimating the IRT outputs; we were able to approximate the IRT outputs nearly perfectly. The coefficients from these regression models can thus be used in the commercial setting to predict individual student performance on each of the relevant DVs that are as rigorous as the more computationally intensive IRT approach. The final computation of student scores was accomplished by multiplying the coefficients by their corresponding raw output scores from iASK provides the student profile of abilities (as described below).

These analyses also provided a way to measure overall performance in decoding and automaticity, to create diagnostic scores for these measures. For these measures, we specifically focused on the output from the vowel GPC model to produce our overall measure of decoding, as vowels show greater variance among struggling middle school readers. We focused on the output from the masked vs. unmasked model to produce our overall measure of automaticity. Note that both of these models included additional factors as well, and so were not completely determined by performance on their focal measures. These models were chosen for both theoretical and data-driven reasons. Theoretically, these models embody the constructs that best map onto overall measures of decoding knowledge and automatic skill use, respectively. Additionally, these models showed quite strong correlations with standard measures of decoding (the mean of the Woodcock-Johnson subscores) and automaticity (the TMSFA). Thus for these models, we included an estimate of overall intercept alongside the other DVs of interest in each multivariate multiple regression model. This provided a set of factors and their coefficients to estimate overall ability in decoding and automaticity. From these estimated overall ability levels, we can assess risk level in each domain by comparing the scores with the standard measures. Because the overall scores are predictive of the standard measures, we can determine cut scores indicating when performance on iASK predicts clinically significant deficits in performance on either the Woodcock-Johnson or the TMSFA.

Table 3: Pearson correlations between the score estimates from the multivariate multiple regression models and those from the IRT models.

Model	DV	Correlation
Decoding: Vowels	Overall	.995
	Short	.989
	Long	.988
	Digraphs	.993
	Diphthong/ R-colored	.985
	Digraph-Irregulars	.974
	Short-Irregulars	.975
Decoding: Consonants	Short (onset)	.970
	Digraph (onset)	.969
	Cluster (onset)	.983
	Short (Offset)	.973
	Digraph (Offset)	.962
	Cluster (Offset)	.964
Generalization: Syllables	1 Syllable	.996
	2 Syllables	.996
	3 Syllables	.995
Generalization: Word/non-word	Word	.997
	Non-word	.997
Masking	Overall	.994
	Masked	.996
	Unmasked	.993

Finally, we can assess what actual student performance looks like in our sample. We can compute scores at two levels for the students. First, we can estimate overall diagnostic student scores in decoding and automaticity; these scores reflect the student's overall ability in each of these skills. Using the coefficients generated from the multivariate multiple regression models, we can arrive at student estimates by multiplying each relevant raw score (of the factors deemed necessary for the estimate by the regression models) by its corresponding coefficient, and then adding these scores to the intercept from the regression models. These scores are in logit space, which is difficult teachers to interpret. To counteract this, we scaled these scores, and assigned diagnostic cut scores based on predicted performance for the WRMT (for decoding) and the TMSFA (for automaticity). Specifically, students who had a scaled score below 200 are predicted to show poor performance in the standard measures (e.g. more than 1 standard deviation below the mean on the WRMT); those between 200 and 300 are predicted to be at risk on the standard measures (e.g. more than half a standard deviation below the mean); and those above 300 are predicted to show little risk on the standard measures.

In our sample, our final computed diagnostic scores were strongly predictive of standard measures of performance. Our overall measure of decoding correlated with the WRMT decoding scores with *r*=.63, and our automaticity measure correlated with the TMSFA with *r*=.68. We thus captured a high degree of variability in standard measures of these constructs, validating these scores as useful assessments of reading ability.

We can also compute students' profiles of performance across the sub-measures. Although these profiles are intended to be used on an individual student basis, we report aggregate performance for the average student here for expository purposes. More detailed comparison of specific student profiles is available in the Teacher's Guide. As in the computation of overall scores, we first multiplied the raw scores from individual factors deemed necessary from the regression models by their corresponding coefficients, and then added these products to the intercept from each regression. Again, these scores are in logit space, so we scaled them to a more interpretable space. We again scaled the scores to a more interpretable range.

These results are displayed in Figure 1. Although there is substantial variability between students on each of these measures, there are several meaningful results that can be drawn from these summary analyses. First, these show that in our sample certain GPC classes and item types are more difficult than others. For example, among the vowel GPC classes, students show particular difficulties with consonant digraphs and clusters, as well as with irregular vowels (especially digraph irregulars) and long vowels. Additionally, students performed better for words than for non-words, and showed a monotonic decline in performance as words increased in number of syllables.

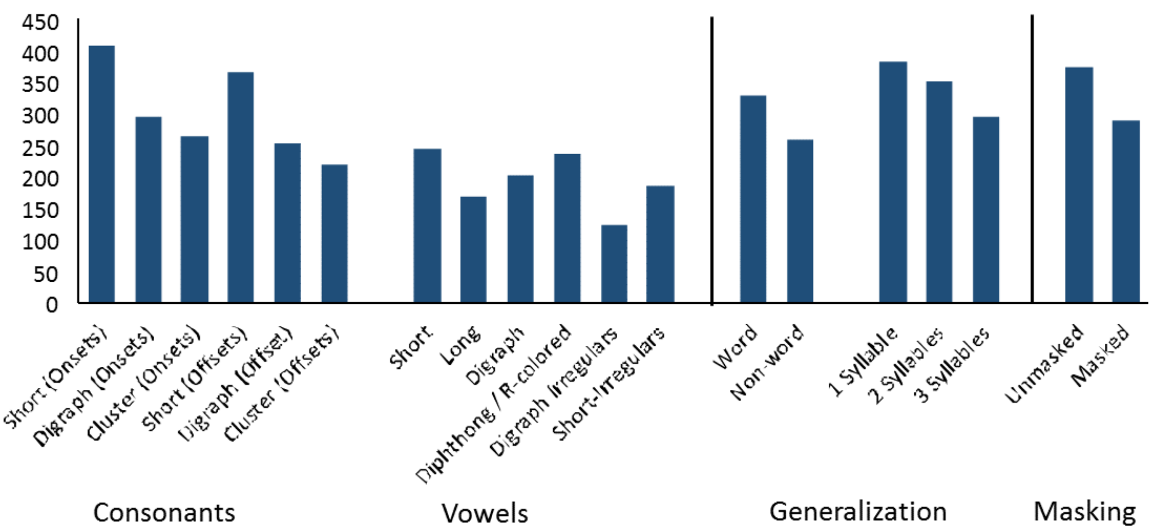

Figure 1: Average student performance on each of the individual factors.

<u>Reliability tests of iASK</u>

Next, we assessed the test-retest reliability of these predicted scores. High reliability is essential for iASK, both to ensure that the resultant scores are being measured accurately, but also to validate iASK as useful across implementations for the same student. Strong test-retest reliability is a prerequisite for use of multiple forms of iASK to assess progress over time. To assess reliability, we used the coefficients estimated from the regression models to compute the estimated participant scores for each ability for both initial data and reliability data for all students who completed both phases. Critically, because the coefficients were generated using only data from the initial phase of data collection, using these coefficients to assess reliability data constitutes a stringent test of our approach. High reliability between initial data and reliability data would indicate both that iASK produces stable estimates of student ability, and that the generated coefficients used to estimate performance are able to capture this stability. We thus measured the test-retest correlation for each DV across all models to ensure that each individual measure is reliable. We compared a given student's estimated performance on each measure at initial data collection to her performance during the reliability phase.

Results showed strong test-retest reliability of all measures (Table 4). Typically, test-retest reliability is considered appropriate with scores above .7, although this value is highly dependent on the measure and sample. In our sample, all reliabilities were above *r*=.73, with most above *r*=.80, signaling that each independent measure was reliable. These scores are particularly noteworthy because different forms of the assessment were used in the initial and reliability sessions; the students showed consistent behavior even when measured across a different set of items. This finding reinforces iASK's ability to measure abilities generally, without being tied to specific lexical items. Additionally, it obviates concerns that findings could be tied to specific item selections. Instead, we are able to produce valid and reliable measures of specific aspects of reading ability using multiple possible forms of our assessment. This ensures that iASK can readily be used multiple times with a particular student, in order to gather formative data throughout intervention. Further, these results demonstrate that the individual measures are each reliable; iASK is thus readily able to extract independent estimates of particular subcomponents of reading ability, and these subcomponents can be measured across the multiple forms of the assessment used.

Table 4: Test-retest reliability scores for all measures. Computed from the estimated scores using coefficients generated by the multivariate multiple regression models.

Model	**DV**	**Correlation**
Decoding: Vowels	Overall	.835
	Short	.835
	Long	.812
	Digraphs	.868
	Diphthong/ R-colored	.863
	Digraph-Irregulars	.739
	Short-Irregulars	.799
Decoding: Consonants	Short (onset)	.758
	Digraph (onset)	.809
	Cluster (onset)	.802
	Short (Offset)	.787
	Digraph (Offset)	.801
	Cluster (Offset)	.781
Generalization: Syllables	1 Syllable	.751
	2 Syllables	.748
	3 Syllables	.766
Generalization: Word/non-word	Word	.854
	Non-word	.860
Masking	Overall	.840
	Masked	.858
	Unmasked	.839

These preceding analyses demonstrate the validity and reliability of iASK's analytic approach. By measuring independent constructs within foundational reading skills, we are able to produce a student's profile of reading abilities, and diagnose deficits in both decoding and automaticity of use. Further, these measures are robust across implementations of the diagnostic; multiple forms can be used across students within the same class, while producing comparable data, and the same student can complete multiple forms of the assessment at different times to assess change in response to training or intervention. These findings confirm that iASK is an effective tool for measuring and tracking middle school reading deficits.

Validating a short-form screener for iASK

The full diagnostic assessment described above proved valid and effective for describing patterns of specific deficits in automatic word recognition among middle school readers. However, this assessment is quite comprehensive and time intensive; some educators may wish to identify which students are likely to benefit from such specificity before assigning students to complete the full assessment. For example, some students may have deficits beyond foundational reading abilities (e.g. vocabulary deficits), and others may have profound deficits that demand more immediate intervention. To this end, we developed a short-form screener version of iASK designed to identify whether students are likely to have foundational word learning deficits, and to quickly gauge the likely severity of such deficits, without the substantial time commitment of the full iASK.

In the full diagnostic assessment, students could demonstrate problems in either basic decoding knowledge or automatic deployment of this knowledge. However, because automatic use is predicated on having the requisite knowledge (i.e. a student can't automatically use knowledge that they don't have), we can look specifically for automaticity deficits as indicative of some form of foundational deficits. The screener does just this, emphasizing deficient automaticity as a rapid measure of whether a student likely has some form of difficulty with foundational reading abilities.

Development of the screener initially built from the data collected for the full diagnostic, by determining how well we could predict full diagnostic outcomes on the basis of subsets of trials. To ensure independence of these predictions, we used data from one phase of data collection (the initial testing phase) to predict performance from another phase (the reliability phase). That is, we sampled a subset of items from students' initial completion of iASK, and investigated how well performance on this subset predicted their full performance when they completed iASK for reliability.

Based on the design and item selection suggested from these analyses, we built a 144-trial screener. We subsequently tested this screener on a novel sample of middle school students, who completed the screener and then the full diagnostic iASK, to ensure that predictive validity remained strong when the screener trials were completed outside the full diagnostic.

Table 5: Tasks included in the final version of the screener.

Task Name	**Blocks**	**Trials per block**
Verify Monosyllabic	2	12
Verify Masked Monosyllabic	2	12
Rhyme Identification	2	12
Rhyme Identification Masked	2	12
Find the Picture Monosyllabic	2	12
Find the Picture Masked Monosyllabic	2	12

Design. The goal of the screener is to achieve reasonable estimates of the overall automaticity scores from the full diagnostic. To this end, the design sought for the screener emphasized: tasks that were most predictive of these full composite scores; and items that were balanced across GPC classes and lexical characteristics to ensure that relevant factors were appropriately assessed. Hierarchical regression analyses demonstrated that the FitB and CtW tasks predicted little variance in the full composite scores beyond the other tasks, so these were not included in the screener. Additionally, the secondary pronunciation GPC classes were less informative, so we emphasized items from the other GPC classes. Finally, the multisyllabic tasks were less essential for predicting full composite scores, so we included only monosyllabic items[2]. These selections helped reduce the overall length of the screener without sacrificing predictive validity. Where applicable, we balanced with respect to word/non-word status. This led to a design with the three remaining tasks (Ver; Rhy; Pic), with both speeded and unspeeded versions of each task. To achieve appropriate balance between the counterbalancing factors and to equally sample each task, we used the design described in Table 5. This led to 144 total trials, divided into 12 equal-sized 12-trial blocks. These blocks were assigned to three locations in the iASK interface, with four blocks of trials at each location.

Item selection required selecting items that met the counterbalancing conditions established for the screener while completing the laid out design. However, because of the variable curricula completed by students during the testing of the full diagnostic, we could not choose a static set of items to investigate across all students. Instead, we assessed how predictive the general design would be across students given variable item selections that maintained the described design. A custom Matlab script randomly chose items that met the design constraints for each student from the initial testing of the full iASK diagnostic; that is, each student from the initial iASK sample had a randomly selected set of items to fill out screener design. This method creates a very conservative measure of the screener's predictability, as between-student item variability could impede performance. However, the high consistency across curricula in the full diagnostic provided confidence that item-level differences were small. In rare cases, exact completion of the intended design was impossible (e.g. a student's full iASK curriculum may have had insufficient items from a particular GPC combination in a task; typically this occurred in the Pic task, because of limitations in picturable items that are likely familiar to middle school students). When such cases arose, replacement items were selected that most closely matched the intended item (changing a single factor, and always choosing an item from the intended word/non-word class).

[2] Note that these factors are informative for aspects of the full profile generated by the diagnostic, and so are necessary in the design of the complete assessment. They proved herein to be unnecessary specifically for predicting the overall automaticity score of the full assessment, and so need not be included in the screener.

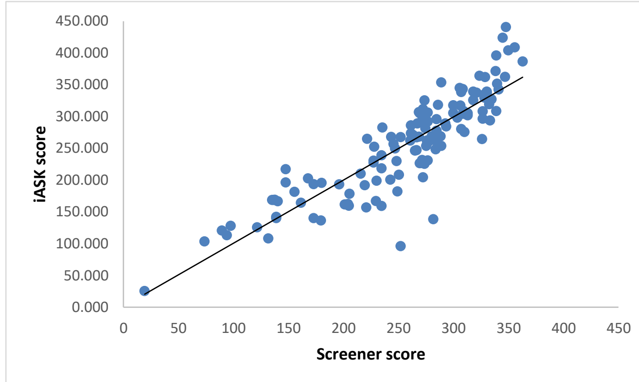

Figure 2: Predicted automaticity scores generated from screener trials, compared against true automaticity scores from the full diagnostic.

This sampling approach provided a simulated screener for each student, based on the subset of trials needed to fulfill the screener design. We then assessed how well the screener could predict their full automaticity score from iASK. In the sample of students used to test the full version of iASK, 128 provided complete data sets in both initial testing and reliability testing. We thus used the data from these 128 students to assess how well the screener scores predict the full iASK score, as well as to determine the appropriate factor weighting for strongest predictive validity. To do so, we conducted a series of regression models, including factors for task, word vs. non-word status, and masked vs. unmasked status from the screener trials drawn from the students' initial completion of iASK, and used these factors to predict the automaticity score from the students' iASK reliability testing. We began with a full model including all task and item factors, and then tested whether simpler models were as effective. For each model, we ran 100 iterations of the item-selection script to for each student to ensure that effects were robust across different random item selections; this analyses showed extremely high consistency across runs. These analyses revealed that only the overall performance on speeded and unspeeded trials were required to effectively predict automaticity performance. We used the average coefficients from the 100 regression models to select the appropriate intercept and scaling for each factor. This provided the formula needed to convert from the screener trials to the predicted automaticity score. As seen in Figure 5, this proved extremely effective in predicting the iASK automaticity scores; the 144 trials in the screener, scaled based on performance in masked and unmasked tasks, elicited r^2=.77.

Next, we chose a particular set of items to serve as the primary screener curriculum for future implementation. To do so, we removed any items that had highly idiosyncratic performance (e.g. items that were anti-correlated with performance or items), and then used another Matlab script to randomly select a subset of the remaining items to complete the design. This was then deployed as a stand-alone screener.

To test this stand-alone screener, we examined a new sample of middle school students from the Cedar Rapids Community School District, who completed both the screener and the full iASK diagnostic. These students had not participated in the previous testing of iASK. These students completed these measures as part of a larger initiative to identify struggling students within the school district. A large number of students from across several schools in the district completed the screener, and those students whose screener scores (i.e. their predicted scores for automaticity on the basis of screener performance) fell below 325 moved on to complete the diagnostic (the remaining students were deemed unlikely to have foundational reading deficits, and so did not devote additional time to this study). This produced a sample of 216 students who completed both the standalone screener and the full diagnostic.

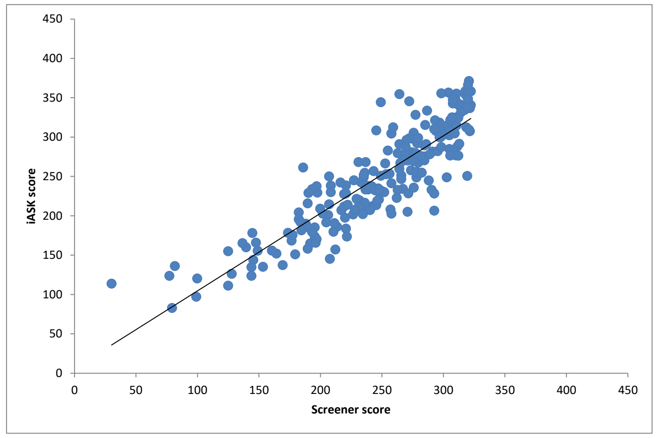

Figure 3: Predicted automaticity scores generated from screener trials, compared against true automaticity scores from the full diagnostic in a novel sample of middle school students.

Within this novel sample, we again found extremely high correspondence between the predicted scores from the screener and the automaticity scores from the full diagnostic (*r2*=.78; Figure 3). This finding confirms that the subset of items used in the screener is extremely indicative of automaticity abilities, even when used as a standalone outside of the full diagnostic. That is, the 144 trials in the screener are quite effective at providing a measure of whether students are likely to suffer from foundational reading deficits. The preceding analyses demonstrate the screener's capacity to predict automaticity scores from the full diagnostic quite well. However, these predictions are imperfect, as expected from using a short-form tool to estimate complex skills. This imperfect predictive validity poses a challenge for the optimal way to convert screener performance to a meaningful representation for educators to understand their students' likely abilities. To this end, we developed an expected range of scores in which we predict a student will fall given their screener performance, rather than a single value. This range accounts for variability in predictive validity, and demonstrates for educators the possible abilities for a given student. The range maps on to the automaticity score for iASK – scores that are within the range below 200 are likely to have major deficits with automaticity; those between 200 and 300 are at risk of automaticity deficits; and those whose range is above 300 are expected to show proficiency with foundational reading skills.

To determine the appropriate range, we tested several possible intervals with both screener validation samples to determine a range that: 1) captures most students' true automaticity scores; and 2) is narrow enough to remain informative. For example, a range of 100 points would capture nearly every student, but would be so wide as to be unhelpful for educators.

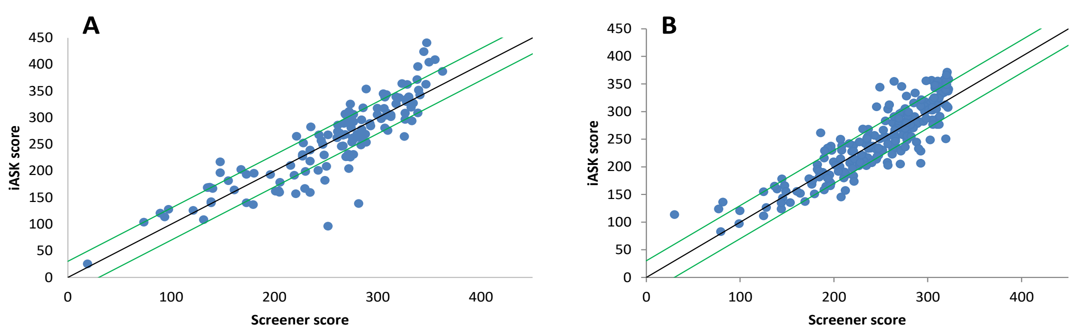

Figure 4: Screener predicted scores compared to automaticity scores from the full iASK diagnostic. Ranges used to interpret screener performance are represented with the green lines. A) Data from the initial sample used to develop the screener design. B) Data from the test sample using the standalone screener.

These analyses settled on a range of 30 above and 30 below the estimated score from the screener (so a student whose screener score estimated an iASK automaticity score of 250 would be classified as falling somewhere between 220 and 280). Using this range is depicted in Figure 4, for both samples of screener testing. As is apparent from the figure, this range captures most students (in the sample from initial testing using a subset of iASK trials to predict full iASK performance, 58% of students fall within the range; in the more stringent testing, using a standalone screener to test a separate full implementation of iASK, 70% of students were within the range).

The results from these analyses demonstrate that the short-form screener is able to capture student automaticity abilities quite effectively. Using results from just 144 trials, that are completed in around 25 minutes, we are able to identify whether a student is likely to exhibit substantial foundational reading deficits. This short-form screener is thus an effective tool to rapidly identify middle school students whose reading is likely hampered by lack of basic reading knowledge or abilities, as opposed to students who may struggle with reading because of higher level difficulties, such as poor vocabulary or comprehension.