

**Validity Testing of the Foundations Learning System Diagnostic  
Assessment Tool for Elementary School Readers**

Keith Apfelbaum  
*University of Iowa*

and

Carolyn Brown  
Jerry Zimmermann  
*Foundations in Learning, Inc.*

***Technical Report Prepared for Foundations in Learning, Iowa City, IA  
December 2019***

*For more information contact:*

Keith Apfelbaum  
Dept. of Psychological and Brain Sciences  
University of Iowa  
G60 Psychological and Brain Sciences Building  
340 Iowa Ave  
Iowa City, IA 52242  
319-335-2444  
[keith-apfelbaum@uiowa.edu](mailto:keith-apfelbaum@uiowa.edu)

## Introduction

The ability to read quickly and efficiently is critical to academic and professional success. However, a majority of American school children fail to master this essential skill during their early education. The National Assessment of Educational Progress (NAEP) consistently finds that a majority of students lack grade-level proficiency in reading ability in middle school. In 2019, this assessment of nationwide academic outcomes found that only 35% of fourth grade students and 34% of eighth grade students were at grade level (US Dept. of Education, 2019). Troublingly, these numbers reflect ongoing trends of failure in reading education, with scores slightly *lower* than in the 2017 NAEP assessment. Additionally, many of the students struggling in these grades exhibit deficits not only in grade-appropriate reading, but also in basic reading skills; only 66% of fourth-graders exceeded threshold for *basic reading abilities*, along with only 73% of eighth-graders. More than a quarter of American students are entering high school without adequate basic reading skills. These skills are a primary focus of elementary school, but many students fail to acquire them during these years.

Deficits in foundational reading abilities as students enter the later years of their education is problematic for several reasons. At the most pressing level, these students will struggle to use reading as a primary means of accessing information across the curriculum. By late elementary school, students begin to pivot from learning to read toward reading to learn – they are presumed to have adequate skills to use text as an information source. Deficits in reading ability curtail this opportunity; students below grade level for reading are unable to effectively draw information from increasingly complicated textbooks, leading to broadening struggles across content areas. In addition, the academic environment for these students is not conducive to remediating these deficits. Few middle school and high school teachers are trained to teach basic reading abilities, as these skills are expected to be mastered by the end of elementary school. Middle and high school students also have little time in their academic schedules for reading remediation; time spent re-training reading skills draws away from time in other content areas, risking increasing deficits for students. Early and effective intervention is thus essential to identify reading deficits before they compound to large-scale academic struggles.

Children need to not just know the rules in the reading system, but also to develop considerable skill in rapidly applying these rules during fluent reading. Skilled readers accomplish the basic processes of reading effortlessly, allowing them to devote cognitive resources to comprehending the content of the text rather than to the process of reading itself. Thus, developing readers need to simultaneously learn the explicit patterns of reading, as well as to effectively apply this knowledge quickly and without conscious effort. This raises a particular challenge for assessing reading development: we can't rely on measures solely of what children know, as we also need to measure their use of this knowledge.

**Summary of the problem:** Current assessments of reading development often assess only what students know, and don't directly measure their ability to use this knowledge flexibly and efficiently. Depending on the provenance of students' deficits, different interventions may be needed. A student that can decode well but does so only with considerable effort may require training targeting speeding their access to knowledge; meanwhile, a poor decoder may require more focus on explicit explanation of the rules of the reading system. Our diagnostic tools must differentiate between these populations, allowing more personalized approaches to reading intervention. Critically, this differentiation of difficulties needs to start early in schooling; rather

than waiting until they reach a point where their poor reading precludes further learning, early identification can help educators catch these students up before it becomes more problematic. Without early and differentiated diagnostics, the startling statistics of reading deficits in middle school will persist.

**The Foundations Learning System approach to reading assessment:** Foundations in Learning previously developed a valid and reliable assessment of decoding knowledge and its use for middle school readers. The validation of this assessment demonstrated that decoding knowledge and its automatic use can be dissociated; specifically, some readers show strong understanding of decoding when given no time constraints, but show substantial difficulty using this knowledge, whereas others show deficits in both decoding and automaticity. Assessing reading difficulties in the componential fashion provides a more valid assessment of students' specific needs.

The Foundations Learning System Diagnostic for middle school students demonstrates diversity of student deficits, both in terms of whether students struggle with decoding knowledge or its use, as well as in the specific aspects of decoding knowledge that students need to improve. This approach overcomes limitations of more unitary models of reading deficits, as it allows customized intervention approaches. Critically, this diagnostic tool has demonstrated predictive validity with multiple measures of different subcomponents of reading, including decoding, as indexed by measures like the Woodcock-Johnson Word Attack and Word ID subscores, and fluency, as indexed by measures like the Texas Middle School Fluency Assessment. Thus the approach used for this diagnostic is able to capture multiple aspects of reading ability using its different constructs of reading ability.

At the core of this approach is the distinction between knowledge and use. Knowledge in this perspective is defined as the ability to recognize the patterns in English orthography, often referred to as the grapheme-phoneme correspondence (GPC) regularities or rules. These patterns dictate how letters map to sounds; skilled readers have deep knowledge of these regularities, including both dominant regularities (e.g. EE makes the long-e sound) and more nuanced regularities (e.g. EA makes the short-e sound when followed by -D, as in *THREAD*, *BREAD* and *HEAD*). Decoding ability signifies the ability to read a word by linking the letters to their sounds; in typical models of reading, this would account for pathways that map orthography to phonology, with meaning accessed through the sounds of words (Coltheart et al., 2001; Seidenberg & McClelland, 1989).

Although successful decoding can allow a reader to read most normal words given enough time, skilled reading requires more than the ability to effortfully map letters onto sounds. Just as a skilled mathematician no longer adds using his fingers, the skilled reader does not need to carefully consider the sound that each letter makes. Instead, this process becomes more rapid and *automatic*. This construct of automaticity signifies the ability to effortlessly access the sound or meaning of a word, without laboriously decoding its constituent parts. The idea of automaticity draws on a long history of research from cognitive science on skill learning, especially in motor learning (e.g., Kellman et al., 2008; Wulf, Shea, & Lewthwaite, 2010; Wulf & Su, 2007). As a skill becomes more fluent and efficient, it demands fewer cognitive resources to be completed. Fluent reading seems to possess the hallmarks of such effortless, automatic processing. Past studies from Foundations in Learning and affiliated researchers have found compelling evidence that the principles of automaticity in skill learning apply in reading (Roembke, Hazeltine, Reed, & McMurray, 2019).

A critical distinction needs to be drawn between *automaticity* and *fluency*. Many extant assessments of reading track a student's level of fluency, often operationalized by things like words correctly read per minute. Automaticity is not synonymous with fluency, but is instead a requisite subcomponent of fluency. Fluent reading requires a constellation of subskills coordinating together to elicit effective productions, including decoding ability, but also efficient use of this decoding ability, processing speed, articulation ability, and others. Diagnosing fluency deficits can help identify students that need additional help, but it lacks specificity of where they fall short. The Foundations Learning System proposes that lack of fluency could arise because of lack of decoding knowledge, or because of lack of automaticity in applying this knowledge. This approach thus aims to measure the level of automaticity as a somewhat independent construct.

This independence is only partial however. Automaticity must be considered with respect to degree of decoding knowledge, not as fully independent. As described in the data analysis presented below, automaticity can be residualized from decoding ability, but it always builds on decoding. This symbiosis occurs because the reader must be automatic *at something*. For example, in skill learning, researchers can assess the automaticity of a golf swing, but this automaticity is meaningless for a person who doesn't know how to pick up a golf club. In reading, the reader needs to have decoding knowledge to be able to deploy it automatically. However, having the knowledge along is not sufficient to signal automaticity. Many readers have strong explicit knowledge, but are ineffective using it; others have only some knowledge, but are quite automatic *with that knowledge*. Thus we can consider the relative degree of automaticity, given a student's level of decoding ability.

The Foundations Learning System thus provides multiple measures that can signal a student's relative level of automaticity, as well as their overall knowledge and effective use thereof. The zone scores signal where a student's level of automaticity falls relative to other students at their current level of decoding ability. The System Scores signal how they put everything together: given their knowledge and their ability to deploy this knowledge, how fluently are they expected to be able to read?

The distinction between knowledge and the ability to automatically deploy that knowledge is critical in planning intervention. Reading education has often emphasized the mastery model, which rests on the assumption that with enough explicit knowledge of a reading skill, it becomes "mastered," and thus known as well as is needed. However, many other domains show clear distinctions between expert-level *knowledge* of a skill and efficient *use* of that knowledge. In these domains, the type of training that best leads to explicit knowledge is often different from the training that best boosts efficient use. For example, Kellman and colleagues have demonstrated across a wide range of domains that expert-level knowledge and fluency with that knowledge require different training (Kellman et al., 2008; Kellman, Massey, & Son, 2010; Romito, Krasne, Kellman, & Dhillon, 2016). In one particularly compelling case, they found that skilled pilots with years of flying experience showed substantial gains in efficiency for flight-related cognitive tasks after a brief, implicit perceptual learning procedure (Kellman & Kaiser, 1994). Building fluency across domains appears to rest on a different set of principles than building knowledge. As such, a reading diagnostic is most effective if it can differentiate between knowledge and use.

### **The Foundations Learning System Elementary School Diagnostic**

The Foundations Learning System already includes a valid and reliable assessment of decoding knowledge and automaticity of use for middle school readers. Earlier ability to differentially diagnosis these deficits in elementary school will provide a way to intervene immediately as

deficits emerge. This early diagnosis could allow students to begin receiving more focused remediation while in an academic environment focused on foundational reading education. To this end, the middle school version of the diagnostic tool was adapted for use with an elementary school population. This report presents the results of initial validation testing of this assessment to determine if 1) an elementary school version of this style of assessment is feasible with an elementary school population; 2) if this assessment shows criterion validity with other measures of reading ability in elementary school students; 3) if these data show a dissociation between knowledge and use in this age group; and 4) if multiple forms of this assessment are equally effective.

The development of this version of the assessment used the middle school version as its basis, and amended tasks and items based on consideration of pilot data with elementary school students and performance by both elementary and middle school students. The elementary school diagnostic assessment consists of a series of tasks and a variety of items that measure student abilities in various ways. This diversity includes items representing a host of GPC regularities, from simple to complex, and uses both real word and nonsense word items. The inclusion of nonsense words allows the assessment to measure decoding ability when memorization of specific words is impossible, and thus provides a purer measure of decoding. Additionally, the GPC regularities are embedded in words of different syllable lengths (from one to three syllables) to measure how students apply their knowledge in simpler and more complex contexts.

The varied tasks in the assessment provide insight into the different aspects of knowledge and use involved in learning to read. Reading fluently involves not just mapping from letters to sound, but also mapping from letters directly to meaning, recognizing comparisons between words, understanding how letters behave in context, and doing each of these things quickly. Previous work validating the Foundations Learning System in middle school, as well as external basic research using this platform, has shown independent predictive value of these diverse tasks. The current assessment thus includes a wide range of tasks to ensure a comprehensive assessment of reading ability.

**Structure of the Assessment.** The assessment included three distinct forms with non-overlapping items. Each of these three forms had identical structures, including the same distribution of tasks, GPC types, number of words and nonsense words, and speeded and unspeeded blocks. Care was taken to attempt to equate difficulty of items in the forms using accuracy data from items in these same tasks completed by middle school students, as well as data from elementary school participants who completed pilot versions of these tasks to determine their feasibility. Additionally, each form located tasks in the same destinations during testing. The intent was to create three independent forms that could be used to assess growth within a student without repeating items.

The complete structure of the assessment consists of 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*”). The letter to change is removed, and students then have to choose the appropriate letter from six 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 base tasks were included to assess different aspects of reading ability. One varied factor was single-syllable vs. multi-syllable to measure ability applying knowledge in different contexts. In single-syllable variants, all items were one syllable. These tasks measure decoding in ideal settings, when context from surrounding syllables is minimized (e.g. for a word like *BABY*, the A is long despite the silent E that is subsumed by the second syllable). In multisyllabic variants, items are evenly distributed between one-, two- and three-syllable items. When appropriate, the location of the targeted response was evenly distributed among the syllable options (i.e. for three-syllable items in FitB, approximately the same number of responses were in the 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> syllable). The types of items in each syllable condition and in each syllable slot for multisyllabic items were also equated (e.g. the complexity of GPC rules for one-syllable items was approximately the same as that for three-syllable words). Finally, where appropriate, the same number of words and nonsense words were included in each syllable condition.

An additional manipulated factor was speeded vs. unspeeded variants to measure how well students could apply their knowledge given no time pressure, compared to how well they did so when under quite intense time pressure. This manipulation has proven uniquely predictive of fluency above and beyond standard decoding measures (Roembke et al., 2019), and thus is taken as a strong index of the automaticity of word recognition. For the speeded tasks, the printed words are shown for an extremely brief period of time (90 msec), followed by a backward visual masker of # symbols. This forces the child to read the item automatically, as the rapid presentation prevents slow, effortful decoding.

Finally, tasks included both word and nonsense word trials, when possible. This was manipulated within a block of trials, such that half of trials were words and half were nonsense words. The use of both types of trials allows the assessment to differentiate between word knowledge and more general decoding knowledge. To decode a nonsense word, a student can't rely on having memorized the specific item, as they item is unlikely to have been encountered before. To minimize any potential learning effects within the assessment, nonsense words each only occurred once in a curriculum.

How these variants were distributed across tasks is detailed in Table 1. Not all variants were included in all tasks, as some were not feasible (e.g. a speeded version of CtW and FitB do not work given the way the tasks are designed; non-word variants of Pic tasks don't make sense); others proved too difficult for this age group in pilot testing (e.g. a speeded version of the Rhyme task led to near floor performance for elementary school students); finally, others proved unpredictable in earlier versions of the middle school assessment, and so were not

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

<b>Task Name</b>	<b>Blocks</b>	<b>Trials per block</b>
FitB Monosyllabic	3	12
FitB Multisyllabic	4	12
CtW	3	12
Verify Monosyllabic	2	12
Verify Speeded Monosyllabic	3	12
Verify Multisyllabic	2	12
Rhyme	2	12
Pic Monosyllabic	3	16
Pic Speeded Monosyllabic	3	16
Pic Multisyllabic	2	12

tested for this version (e.g. speeded multisyllabic tasks proved to offer no predictive validity that wasn't garnered by other tasks).

Blocks of tasks were instantiated as “levels” that were distributed across seven locations on a cartoon island. Each destination included three or four tasks. Similar variants of tasks were placed in different destinations (e.g. no destination included more than one block of FitB; masked variants were spread approximately evenly across destinations). Students could choose which destination to complete, but once they began a destination, they were required to complete all tasks therein before proceeding to a new destination. Within the destination, the student chose the order to complete the tasks.

***Items.*** In general, each task was counterbalanced for the targeted GPC classes (Table 2). These classes were approximately evenly balanced between and within each task by word vs. non-word status (when applicable – nonsense words did not include secondary pronunciations, as these are unlikely). In some cases, limitations in available appropriate items made perfect counterbalancing impossible. In these cases, balancing was maximized, and any imbalances were equated across the three forms of the assessment. Balancing primarily focused on the targeted forms (e.g. the targeted letter in FitB trials); however, the surrounding context was also balanced as best as possible (e.g. for trials targeting vowels, the complexity of the consonants in those items was controlled across the vowel GPC classes). The choice of the targeted form superseded balancing the context when conflict arose.

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. Care was taken to ensure that these variants had extremely similar balances, and that the overall difficulty of items included in the variants was matched as best as possible. Although the statistical model used to assess the data can accommodate item-level differences, maintaining approximate equivalence of item difficulty makes interpretation of raw performance between tasks more feasible. Additionally, 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 an elementary school student is expected to know). In these cases, items were chosen that were as close to the target characteristics as possible.

In tasks that require targets within items, such as FitB, the location of targets was selected to optimize the informativeness of trials, while also balancing as best as possible. For example, in early reading development, vowels tend to be the most problematic GPC classes. As such, FitB tasks included twice as many vowel trials as consonant trials. Within the vowel trials, vowel GPC classes were balanced. Within consonant trials, half of items focused on onset consonants, and half on offset consonants. The type of consonant GPC class was balanced for both onset and offset consonants.

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

<b>GPC Class</b>	<b>Description</b>	<b>Examples</b>
CS1	Simple consonants at the start of a syllable. These are single-letter consonants that produce a single sound.	<b>Cat; moon; building</b>
CDG1	Consonant digraphs at the start of a syllable. These are multi-letter consonants that make a single sound (e.g. SH, CH, WR).	<b>Shepherd; knife; whale</b>
CCL1	Consonant clusters at the start of a syllable. These are multi-letter consonants that make multiple sounds (e.g. CL, BR, SPR).	<b>Club; broom; splitting</b>
CS2	Simple consonants at the end of a syllable. These are single-letter consonants that produce a single sound.	<b>Boot; crab; random</b>
CDG2	Consonant digraphs at the end of a syllable. These are multi-letter consonants that make a single sound (e.g. SH, CK, LL).	<b>Doorbell; duck; crash</b>
CCL2	Consonant clusters at the end of a syllable. These are multi-letter consonants that make multiple sounds (e.g. SK, LD, TCH).	<b>Task; patchwork; gift</b>
SH	Short vowels. Single-letter vowels that produce their dominant pronunciation.	<b>Cat; fetch; puddle</b>
LO	Long vowels. Vowels with the long-vowel pronunciation driven by a syllable-final silent E.	<b>Bite; parade; bone</b>
DG	Digraph vowels. Two-letter vowel combinations that produce their dominant pronunciation (e.g. EA, OA, AI).	<b>Bait; cleaner; root</b>
DR	Diphthong and r-colored vowels. Letter combinations that create either diphthongs or r-colored vowel sounds.	<b>Bark; bound; boiler</b>
DS	Digraph secondaries. Multi-letter vowel combinations that create either secondary pronunciations or exceptions. For example, words when EA makes a short-vowel sound.	<b>Head; book; four</b>
SS	Short secondaries. Single-letter vowels that create either secondary pronunciations or	<b>Ball; blind; full</b>

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). Items were pseudo-randomly assigned so

that predicted difficulty was approximately equal across tasks. We generated three full item sets (curricula), with no overlap between sets. The three unique curricula can determine reliability of our measures independent of specific item choices, and they can allow multiple implementations of the assessment on a single student, providing a means of progress monitoring. For data collection, students were randomly assigned to one of the three curricula, with care taken to balance gender, grade and school.

**Participants.** Participants were drawn from two separate samples, each run in the spring semester of the 2018-2019 school year. Most students were drawn from schools in the Distinctive Schools charter school system in the Chicago area. From this sample, second, third and fourth graders who had completed external measurements using MAP Growth were selected for consideration in the validation study. A total of 601 students from this sample completed the entire Foundations Learning System Diagnostic: 206 second graders, 209 third graders, and 186 fourth graders. An additional 80 fourth graders from the West Des Moines school district completed the assessment as well. This further sample was recruited to provide an ability to measure fluency in the fourth-grade age group. Distinctive School provided MAP Fluency scores for second and third grade students, but fluency data for fourth graders were unavailable. The West Des Moines sample included other external measures of fluency to gauge the scaling used from the younger ages in Distinctive Schools. These students were randomly assigned to the three forms of the assessment, with care taken to approximately balance by grade and gender.

**Procedure.** Participants completed the assessment in their schools during the spring semester of the 2018-2019 school year. All participants were enrolled in the assessment and provided with enrollment IDs and passwords by Foundations in Learning. For participants in the Distinctive Schools cohort, research assistants from Foundations in Learning were on hand to assist with initial login and at least the first day of testing in each school; when needed, these RAs returned to the schools for further assistance. In most cases, teachers took over administration after the first day. Data collection in the West Des Moines School District was conducted by district staff, who have substantial experience with Foundations in Learning software from other implementations in the district.

Students completed the assessment in installments of 20-25 minute sessions over successive school days. The program is delivered entirely via a web interface, and each student completed the tasks using a laptop and wearing headphones. Each student completed a single session per day, and most students completed the entire assessment in three sessions. When needed due to absences or slow completion, students completed additional sessions to fully complete the assessment. Students logged in using a unique username, and their progress was automatically saved each day. When students reached the end of their time limit, they were allowed to complete their current task, and were then automatically logged out. When they next logged in, the new session began wherever they left off at the end of their previous session.

***Benchmark Measures.*** The school districts provided standardized assessment data for the students enrolled in this validation study for use as benchmark measures. These assessment data were collected in the normal course of school activities, and their implementation was not overseen by Foundations in Learning. These benchmark measures were all collected during the later part of the spring semester, within 6 weeks of the administration of the Foundations Learning System Diagnostic, to ensure that the assessments were measuring students at around the same level of ability. All students in Distinctive Schools completed the MAP Growth assessments (MAP Growth: Reading 2-5 CCSS 2010 V3). MAP Growth measures grade-level student performance in math and reading, including multiple valid measures of subskills of each. This measure has been validated for all included grades. We focused on percentile ranking of

students in reading within their grade as the benchmark measure for decoding. Note that the West Des Moines sample did not complete MAP Growth, and so was not included in the validation of decoding measures.

MAP Growth is predictive of other decoding measures, like the Woodcock-Johnson, but does not provide a dedicated measure of reading fluency (which automaticity is expected to be predictive of). However, a more recent measure from the same developers, MAP Fluency, directly assesses students' passage reading fluency levels. This assessment uses grade-level-normed passages to determine individual students' levels of fluency. At the time of the current validation study, MAP Fluency only had grade-normed passages up to grade 3. Students at Distinctive Schools in second and third grade thus had MAP Fluency data; percentile scores from this assessment were used to validate the automaticity and system scores for the Foundations Assessment. No valid measures of fluency were available for the fourth graders in this sample. As such, a cohort of fourth graders from the West Des Moines school district were recruited; this district collects FastBridge aReading scores, which provides a broadband measure of reading ability that subsumes numerous components of reading skills. This measure was used with this set of fourth graders to confirm that the scaling developed from the younger grades at Distinctive Schools were highly predictive of this other omnibus measure of reading ability in fourth graders.

**Results.** Data were analyzed separately for the different constructs of interest: decoding (and the decoding sub-scores), automaticity, and the Foundations System Score. For each factor, we used an iterative process to identify the appropriate scaling, and then to identify an effective means of predicting the scaled scores in a tractable timeframe.

The analysis plan mirrored that used for the middle school diagnostic assessment to ensure continuity of measurement. For each construct, we initially considered which factors were predictive within the middle school homologue for the construct. Where possible, we maintained these factors in the model for elementary school; when not (e.g. when a factor was not included in the elementary school sample), models were adjusted to include as many viable factors as possible.

After identifying the relevant factors for each model, we conducted item-response theory (IRT) analyses using these factors to determine how to best weight each factor to predict performance. For these IRT models, trial accuracy served as the DV. Fixed effects included all potential variables that were not of interest, and random effects included both targeted variables of interest and the variables not of interest. Excluding fixed effects for variables of interest ensured that participant-specific random effects accounted for all relevant variance in the variable of interest, without allowing the fixed effect to bias these random effects. Forming participant predictions entirely from random effects for the variable of interest provides a more precise estimate of an 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 obscure the effect of interest, as the random effect would subsume overall variance that is better accounted for by the participant effects of word vs. non-word.

The IRT models provide estimates of participant abilities in the various constructs of interest that accounts for idiosyncratic variance from factors like overall difficulty of different tasks or item-specific effects. This form of IRT modeling is a gold-standard approach to estimating student performance (and these specific forms of IRT models proved extremely effective in predicting external outcome measures in previous analyses of the middle school version of this diagnostic, and show predictive validity again in the current analyses). These analyses are thus a useful tool for predicting precise levels of student reading ability across the constructs of interest, and for producing profiles of performance on these abilities.

Unfortunately, IRT analyses are computationally intensive, and thus not viable for producing immediate reporting on student performance. To ensure timely delivery of scores, a more computationally tractable approach is needed. We conducted multivariate multiple regression analyses using the scaled scores from the IRT models as the DV, and overall performance in factors that may be relevant to the model as the predictors. This allows us to identify which specific factors are meaningful for predicting the outcomes for that model, and to produce a linear scaling of scores to produce estimates of the IRT scores. Multivariate regression analyses produce coefficients for each included factor, signifying which are most predictive. As our goal is to produce a highly precise estimate of student performance, and given our fairly large sample, we relied on a conservative significance criterion to determine which factors to include in our final scaling ( $p < 1 * 10^{-10}$ ). The resultant linear scaling from these multivariate provides a computationally simple yet precise approximation of the IRT scores.

To summarize: each outcome construct was specified based on the factors that were predicted relevant for the middle school version of that construct (as possible); these factors were used to conduct IRT analyses to determine estimates of overall student abilities on each factor; and the outcomes of these models were used as DVs in multivariate multiple regression models to develop simple linear scalings of student performance on different factors that could be used to estimate these values. These linear scalings then form the basis of the scaled scores provided as the outcomes for the Foundations Learning System Diagnostic, as well as the scores used to validate the assessment against benchmark measures. We now describe the outcomes of these models, and detail any variations used to make predictions for each of these models.

*Decoding.* The decoding models had two primary goals: 1) Estimate each student's *overall decoding ability*; and 2) to estimate their specific ability levels in decoding subcomponents (each type of GPC class, vowels, and consonants). We constructed separate models for decoding (overall and GPC classes) and decoding (vowels vs. consonants) using the relevant factors derived for the middle school version of the diagnostic. For these models, we included data from all students to generate the most precise estimates possible given our entire sample. Note that derivation of our scale scores does not scale based on students' external assessment scores, but only from bottom-up data gathered within the diagnostic, so including students without benchmark measures is appropriate.

The multivariate models used to predict the decoding scores did so extremely effectively. For each DV, correlations of the IRT output scores with the predicted values from the regression models were very high (all  $r > .97$ ). This demonstrates that the simpler linear scaling of participant performance on the relevant factors was able to predict the IRT outputs nearly perfectly; these scaled scores are thus viable means of predicting student performance levels within the diagnostic. These scaled scores are the basis for further analyses.

These scaled scores also form the basis of validation against benchmark measures. To do so, students that had completed the MAP Growth assessment and the diagnostic were identified. This comprised a sample of 600 students from Distinctive Schools for whom both measures were

available (206 2<sup>nd</sup> graders, 209 3<sup>rd</sup> graders and 185 4<sup>th</sup> graders). This analysis demonstrated moderate correlations of the overall decoding score and students' MAP Growth reading percentiles ( $r=.38$ ), demonstrating that these decoding scores are predicting aspects of MAP Growth. Note that this correlation emerges despite models not being optimized to predict this specific assessment; model structure was determined based on what was most predictive in the middle school version of the assessment to ensure good continuity in the assessment across grades, yet moderate correlations with this external measure still emerged.

*System Score.* In addition to the decoding score and subscores, the Foundations Learning System provides a *System Score*, which serves as an index of the combined decoding and automaticity ability of the student. That is, this score provides a more comprehensive view of the students' ability that includes the multiple underlying components of reading. This score should index reading fluency to some extent, as it includes two of the primary facets that govern how fluently a student can read text (their knowledge of decoding regularities and their skill in efficiently deploying this knowledge).

This score was computed following the same approach as the decoding score: appropriate factors were initially drawn from the middle school models, and these were used to develop the IRT models. The factors included for the System Score were broader, including, for example, both speeded and unspeded versions of tasks. Next, regression models were used to develop linear scalings of

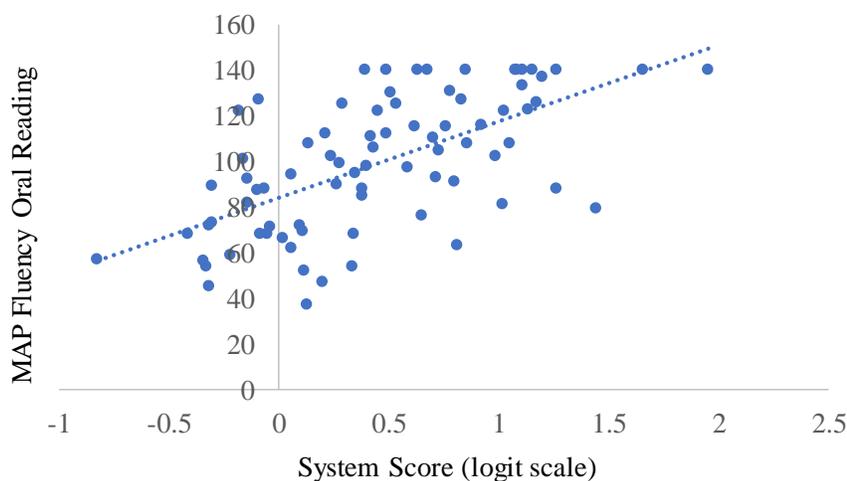


Figure 1: Correlation between System Scores and MAP Fluency Oral Reading scores

factors that proved predictive of the IRT scores. These final scaled scores were compared against available measures of fluency (MAP Fluency) or overall reading ability (aReading), as available to determine how well the System Score predicted external measures reading assessments. These scores were available for a subset of the 2<sup>nd</sup> and 3<sup>rd</sup> grade students at Distinctive Schools ( $n=80$ ; 46 2<sup>nd</sup> graders, 34 3<sup>rd</sup> graders). The System Score proved highly predictive of the MAP Fluency Oral Reading Rate ( $r=.63$ ; Figure 1).

This high correlation signifies that the Foundations System Score is highly predictive of external measures of fluency for 2<sup>nd</sup> and 3<sup>rd</sup> graders. However, as the elementary school version of the Foundations Learning System Diagnostic is intended to cover through 4<sup>th</sup> grade, confirmation that the System Score predicts performance for this older age group is essential. Unfortunately, at the time of testing no 4<sup>th</sup> grade version of MAP Fluency was available, and no other fluency measures were available for the Distinctive Schools sample. However, a separate sample from West Des Moines completed the diagnostic as well as FastBridge aReading, which offers an omnibus measure of reading ability for this age group. From this sample, data for both aReading and the Foundations Learning System Diagnostic were available for 77 students. This sample showed similarly high correlations between the System Score and students' aReading

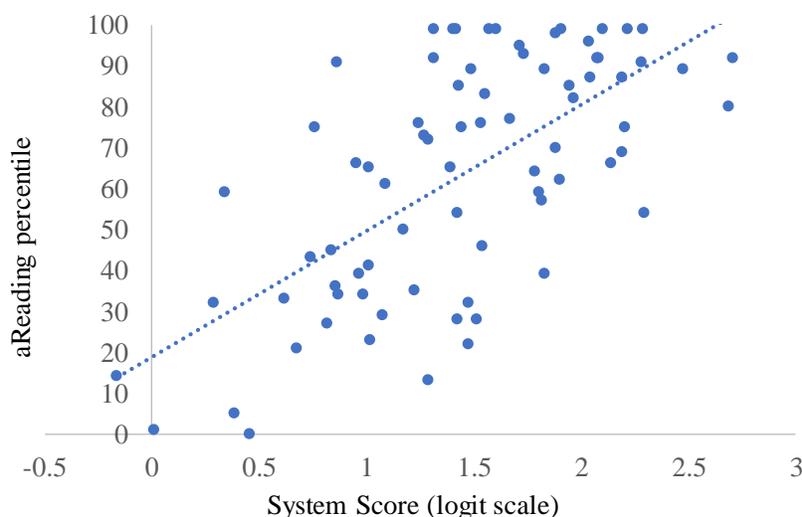


Figure 2: Correlation between System Scores and aReading percentiles

percentile ranks ( $r=.66$ ; Figure 2), confirming that the System Score is predictive for this additional age group, and when using a different external assessment. The System Score thus is a valid estimate of a student’s skilled reading ability.

*Automaticity.*

Whereas the previous analyses demonstrated students’ decoding knowledge and their overall reading ability, respectively, a critical additional measure is students’ relative ability to

effectively deploy their knowledge: how *automatic* they are in decoding. Unlike the preceding analyses, this construct is inherently dependent on the level of knowledge a student has; a student can’t be automatic with knowledge they don’t have. To accommodate this dependence, the analysis of automaticity takes a critically different approach: each student’s predicted decoding scaled score is included as a covariate for the IRT models of automaticity. This inclusion ensures that automaticity is considered *relative* to a student’s decoding ability; the score thus estimates how well they are using their knowledge, given the amount of knowledge they have. Thus a student might be a relatively good user of their decoding knowledge despite low overall levels of knowledge, or a poor user despite high levels of knowledge (and so forth). This distinction is critical, as the different constellations of skills may demand different forms of intervention – in the former example above, the student may need focused explicit training on additional knowledge, whereas the latter example student may need more implicit training to become more effective in deploying her knowledge. These models again took the same overall form as the middle school version. Critically, these models have obligatory inclusion of masked performance to ensure that the automaticity score accounts for differences in performance when time pressure is applied. The outcome for this analysis is a form of residualized performance score: given the student’s level of knowledge, how automatic are they compared to where they would be expected to be? Raw automaticity scores below 0 signify that the student demonstrates worse automaticity than peers with similar decoding knowledge; raw automaticity scores above 0 signify that the student demonstrates better automaticity than their similar-knowledge peers.

The raw automaticity scores are difficult to interpret independent of the student’s decoding score – strong automaticity for a poor decoder likely means something very different from strong automaticity for a strong decoder. The same automaticity score could signal very different abilities; a positive score means only that a student is more automatic *than would be expected given their decoding ability*. As such, automaticity scores are recombined with decoding scores to produce a “zone score.” These zones help identify a student’s likely intervention needs, given their automaticity relative to their decoding ability. Specifically, for students with low decoding ability, the focus of intervention should likely be on increasing knowledge, no matter how automatic students are with the little knowledge they already have.

However, as level of decoding knowledge increases, the degree of automaticity becomes more critical; a student with high decoding knowledge and low automaticity needs intervention that specifically targets their ability to use their knowledge more efficiently.

These zone scores first consider a student’s level of decoding, and then integrate their automaticity. Students in the high-risk decoding group (below a scaled decoding score of 200, as described below) will fall into Zone 1 or Zone 2. Zone 1 comprises the high-risk decoders who also show relatively poor automaticity (scores below 0); Zone 2 comprises high-risk decoders with relatively strong automaticity (scores at or above 0). Again, both of these groups may primarily need intervention focused on improving decoding, despite their differential automaticity. For students in the some-risk decoding category (scaled decoding scores between 200 and 300), their automaticity determines whether they fall in Zone 3 or Zone 4. Zone 3 comprises some-risk decoders who are also below expectations for automaticity; Zone 4 comprises some-risk decoders who are relatively strong with automaticity.

For strong decoders, an additional criterion is included. If students show proficient decoding (a scaled decoding score above 300) and a high System Score (a scaled System Score above 260), they fall into Zone 6, signifying that their foundational skills are likely sufficient; although some of these students may exhibit raw automaticity scores below 0, their overall performance is strong enough to suggest that these differences are less meaningful. These students are strong readers, even if they are not quite as efficient as their similar-knowledge peers, and they likely do not need time devoted to automaticity training. However, if their decoding score is above 300 but their System Score falls below 260, this signifies that despite proficient decoding, the student could benefit from additional automaticity training. These students thus fall into Zone 5.

*Scaling.* The scores output by the preceding analyses are scaled in logit space, meaning that average performance for a given construct would fall around 0, with scores ranging above and below. These scores are not intuitive to interpret, so the scores are linearly scaled to a more intuitive scale. Scores were scaled to offer meaningful diagnostic cutoffs. Specifically, the decoding scores were scaled relative to the MAP Growth scores, such that a scaled Foundations decoding score below 200 corresponds to a predicted MAP Growth percentile below 40, and a decoding score below 300 corresponds to a predicted MAP Growth percentile below 50. Thus scores above 300 are expected to be above the 50<sup>th</sup> percentile as measured by MAP Growth. These scaled scores are capped at 0 at the low end and 600 at the high end. Although mathematically feasible for students to score beyond these extremes, interpretation of differences beyond them are likely unhelpful; a score below 0 effectively means the student is at extreme risk because of poor decoding, and a score above 600 means that the student’s decoding is well into the proficient range.

The System Score was similarly scaled between 0 and 600. In this case, a MAP Fluency Oral Reading Score of 85 was used to demarcate the high-risk grouping (a System Score of 200), and a MAP Fluency Oral Reading Score of 100 signified a System Score of 300. Again, these scaled scores were capped at 0 and 600.

***Reliability.*** The present validation study did not include direct reliability measures, as time constraints precluded having students complete multiple forms of the assessment. Because of the methods used to compute scaled scores, traditional split-half approaches to reliability are not feasible; the assessment’s design is extremely streamlined, so split-half measurement approaches would create

Table 3: Average percent correct responses in the different Forms of the Foundations Learning System Diagnostic by grade.

Grade	Form A	Form B	Form C
3	63.1%	58.5%	62.0%
4	70.8%	69.6%	66.4%
5	67.1%	73.3%	74.1%
<b>Overall</b>	<b>67.0%</b>	<b>67.3%</b>	<b>67.5%</b>

empty cells which would prevent computation of accurate scaled scores. However, informal approaches to investigating whether the multiple forms are comparable show promise. The three forms showed extremely similar overall performance across students (Form A: 67.0% correct; Form B: 67.3%; Form C: 67.5%), and performance was generally consistent across forms within a grade (Table 3). The predicted decoding scores also show general consistency, though with slightly lower scores for Form A (Form A: mean decoding = 307; Form B: 321; Form C: 332). Finally, correlations with MAP Growth percentiles were similar for Forms B and C, but were substantially lower for Form A (Form A:  $r=.24$ ; Form B:  $r=.42$ ; Form C:  $r=.47$ ). Similar results emerged for average System Scores (Form A: average System Score = 303; Form B: 332; Form C: 316). Correlations of System Scores with MAP Fluency suggested that Form C was substantially more predictive (Form A:  $r=.54$ ; Form B:  $r=.54$ ; Form C:  $r=.76$ ). However note that these latter values draw from quite small samples (29 students completed Form A and had available MAP Fluency data; 25 students did Form B and had MAP Fluency data; 26 students did Form C and had MAP Fluency data). As such, caution should be used before concluding that Form C is reliably more effective than the other forms.

These comparisons between forms suggest that there is general comparability between the forms, and each is fairly predictive of the outcome measures. However, these data fall short of a formal reliability measure. Future analyses directly assessing the test-retest reliability of these forms within students will help determine the degree to which they are truly reliable measures of the underlying constructs.