WATERFORD ASSESSMENT OF CORE SKILLS
A Computerized Adaptive Reading Test for Pre-K through 2nd Grade

Haya Shamir, Erin Phinney Johnson and Kimberly Brown
WaterfordResearch Institute, 55 West 900 South, Salt Lake City, UT, U.S.A.

Keywords: Computerized Adaptive Test, Reading Assessment.

Abstract: The Waterford Assessment of Core Skills (WACS) is a new computerized adaptive test of early literacy for students in Kindergarten through 2nd grade. WACS includes assessments in letter recognition, letter sound and initial sound recognition, blending, segmenting, reading real and non-words, reading comprehension, listening comprehension, and vocabulary. A CAT for this age group will be highly beneficial by allowing whole classes to be tested together without additional personnel, by assessing a large number of content areas in reduced time and with fewer questions than a standard paper and pencil test, by producing immediate and accurate score reports, and by engaging students with animations during the test. Reliability and validity analyses indicate that the test is internally coherent and that the subtests correlate well with other reading tests used with this age group, including DIBELS, IRI, ITBS and TPRI.

1 INTRODUCTION

For the last 33 years the Waterford Research Institute has strived to develop high-quality educational models and programs to enable all children to receive the finest education possible. Over time it has become evident that the ability to easily assess student skills in the youngest grade school group, when students are most responsive to intervention, is lacking. Many current assessments available for this age group, Kindergarten through 2nd grade, require one-on-one administration (e.g., DIBELS; Good & Kaminski, 2003), which may result in a great deal of error variance due to differing administration techniques by individual testers or differences in scoring ambiguous answers. In addition, tests that do not require one-on-one administration are limited in the scope of what they can cover and they risk introducing too much variance from fidgety 6-year-olds (e.g., ITBS; Hoover, et al., 2003). A computerized test, such as the Waterford Assessment of Core Skills could provide consistent and efficient test administration by removing the need for a human test administrator and providing an engaging testing environment.

2 METHODS

The Waterford Assessments of Core Skills (WACS) is a web-based adaptive literacy test for pre-kindergarten to second grade students. The new test, soon to be available to schools and home users, assesses early literacy skills including letter recognition, letter sound and initial sound recognition, blending, segmenting, the reading of real words, non-words, and sight words, and comprehension of paragraph-level text. WACS also assesses early language skills including vocabulary and listening comprehension. As an adaptive test, WACS can assess a large number of content areas in reduced time and with fewer questions than a standard paper and pencil test. In addition, computerized adaptive tests (CATs) may reduce frustration for lower performing students and boredom for higher performing students.

2.1 Design

The award winning product design team at Waterford (Software and Information Industry Association, 2008) have created an engaging test that students actually enjoy taking. Throughout the test students are guided by a groundhog named Wyatt (see Figure 1) who is asking for their help on a number of tasks they have to do together. At the

Shamir H., Johnson E. and Brown K. (2009).
WATERFORD ASSESSMENT OF CORE SKILLS - A Computerized Adaptive Reading Test for Pre-K through 2nd Grade.
In *Proceedings of the First International Conference on Computer Supported Education*, pages 24-30

end of each section and at the end of the assessment the students receive a non-judgmental reward screen that serves as a short mental break of dancing characters, with a new character appearing after each skill has been completed. When students have fully completed WACS, Wyatt presents them with a deputy badge as a reward for finishing the assessment.

WACS may be used by home users as well as school users, allowing home-schooling parents access to assessment tools similar to those used in the public or private schools. Home users will be able to complete the test by streaming the required media, while schools will be expected to download the related media to the computers in the classrooms or computer labs.

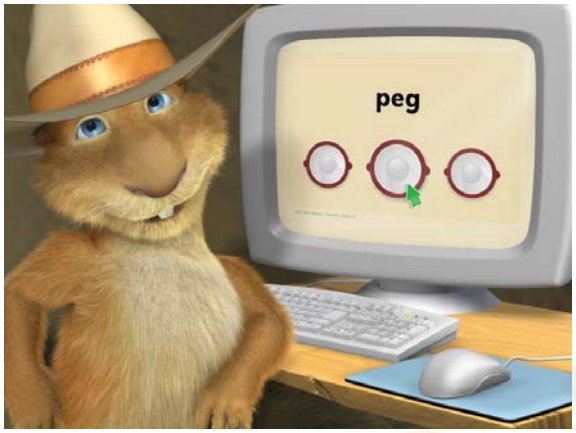

Figure 1: Wyatt, demonstrating the Sight Words Assessment.

2.1.1 Item Design

All test items are presented visually and aurally but do not require the child to speak the answer. For Letter Sound, Real Words, Sight Words, and Nonwords, the letter/word appears on the screen with three speakers underneath (see Figure 1). Each speaker says a different letter/word name. The student must click on the correct speaker to match the word or letter that is on the screen.

Vocabulary differs slightly from this arrangement with a sentence missing one word written at the top of the screen. The sentence is read to the child and the child must pick a word from the speakers that best completes the sentence.

In the case of Letter Recognition, Initial Sound and Blending there is a single speaker or picture at the top of the screen and three pictures or letters at the bottom of the screen (see Figure 2 for example). The speaker/picture emits a sound or a series of sounds and the student must select the picture below that begins with that same sound, matches the series of sounds, or select the letter that matches the letter name from the speaker.

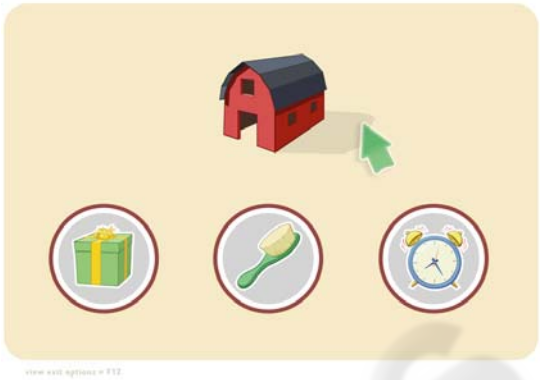

Figure 2: Example of the Initial Sound Assessment. The answer is selected by clicking on the chosen letter with the cursor, the green arrow.

For Reading and Listening Comprehension the student is given a passage to read or listen to. When finished, the child is presented with a question about the passage followed by three possible answers. In Listening Comprehension the questions and answers are presented aurally. Each reading or listening passage includes four questions of varying difficulty. Each child receives three passages depending on skill level.

The final task, Segmenting, differs from all of the other tasks. Here, a picture is presented to the student and he/she must move a series of blocks representing the sounds into the correct order for the word associated with the picture.

For all skills, the computer introduces the question including the correct answer and the distracters. The student can use the mouse to roll over the question or the answer options to hear the instructions again.

2.1.2 Sequence Design

Importantly, all children do not receive all assessments. WACS includes eleven different assessments, a subset of which is given to students depending on their grade level and performance. Limiting the number of assessments completed for each child continues to reduce time required to test the student and allows for a more pinpointed report. It is assumed that students who are advanced in reading do not need to be tested on pre-reading skills such as letter recognition. However, students from advanced grades may receive basic skills if they fail

to complete advanced skills at their grade level. On the other hand, advanced students at lower grades may receive more advanced skills if they prove to be competent at the more basic skills.

2.1.3 Report Design

One of the greatest benefits of computerized testing for students is the ability to receive scores immediately after testing has completed. WACS has been designed to provide reports about individual test takers, as well as class, school, and district level reporting. The reports indicate the child's grade level for each of the completed assessments as well as detailed information about what was actually being assessed and ways that any problem areas could be addressed in the home or in the classroom. The past three test results for the students are also generated on the report, allowing parents and teachers to compare changes over time in relevant assessments.

2.2 IRT Analysis

In September, 2007, 8,800 students in Utah, Idaho, Nevada, California, New York, Texas, North Carolina, and Florida completed the first round of testing with WACS. This first group was given a random sample of questions from each assessment, all questions representing varied expected difficulty levels. The sample of students from twenty six schools was representative of US socio-economic status, ethnicity, geographic location, and type of school, based on information obtained from the US 2002 census.

Based on the responses, difficulty values for these 2,680 items were calibrated using the Rasch model analysis for item response theory. Results revealed 131 items with an outfit mean square greater than 1.7, indicating high error variance in the item. These items were excluded from the test. An additional 131 items with outfit mean square less than .5 were excluded, since items with outfit mean square smaller than .5 are considered less productive to the measure. Subsequent differential item functioning (DIF) analysis revealed 21 items that had a gender bias. These items were removed for content review. Item difficulty was then calculated for the remaining items.

Utilizing IRT analysis on test items, the adaptive nature of the test allows a student's response to determine the next set of items. For example, if a student fails to answer a question correctly within a skill area, the next question he receives will be less difficult. If the student answers that second question correctly, the next question is harder, but not as hard as the previously missed question. In this way, a computerized adaptive test identifies the student's skill level in a particular area. Because WACS can test up to eleven different areas, detailed information about the student's abilities are subsequently available to teachers and parents.

2.3 Validity

Validity is the argument that a specific test score interpretation or use is valid. In other words, a test is valid when it does what it is supposed to do. There are three major categories of validity: those associated with content, criterion, and construct.

2.3.1 Content Validity

In order to establish content validity, this paper discusses the reasons for the test design and content as well as the association between the given test and state standards or curricula. First, content experts investigated the most important skills for pre-kindergarten through 2nd grade students and established guidelines for writing items based on published research. The areas covered included acquisition of letter names and sounds (Adams, 1994; Evans, 2005), early phoneme awareness (Wilson, 1996), sight word reading (Carroll, 1971; Wilson, 1996), real and non-word reading (Wilson, 1996; Ganske, 2000), vocabulary (Stemach & Williams, 1988; Beck, McKeown, and Kucan, 2002), and reading and listening comprehension (Snow, Burns, & Griffin, 1998). In addition, all comprehension passages and questions were written by professional writers, reviewed by content experts, and edited by writing experts. Reading comprehension passages were Lexile certified for their grade levels.

Upon completion, items from all of the subtests were reviewed by additional content experts and sent to Marilyn Jager Adams, an external content expert, for review. After IRT testing, analysis was conducted to insure that item difficulty, as determined statistically by IRT analysis, correlated with the item difficulty as determined by the content experts.

In addition to creating items based on researched concepts, a valid reading test should also cover standards accepted by the states for reading and language development. Thus, state standards were examined and correlated with WACS skills and items. With the exception of Iowa (which did not list

standards below grade three), a minimum of three, and a maximum of eleven, WACS assessed skills were also listed as state education standards for PreKindergarten through grade two.

2.3.2 Criterion-related Validity

The effectiveness of a test in predicting performance on a related task can be measured by assessing performance on two tests at the same point (concurrent validity) or at two different time points (predictive validity). To assess concurrent validity WACS was administered to students nationwide in September and October. Student performance was then compared to performance on the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), the Idaho Reading Inventory (IRI), the Iowa Test of Basic Skills (ITBS), and the Texas Primary Reading Inventory (TPRI). Additional test data for the Stanford Achievement Test (SAT 10) and ITBS will be collected in April, 2009.

Data for predictive validity will also be collected in the spring. WACS will be administered to the same students from the fall testing. Those students are also completing a spring round of testing for DIBELS, IRI, or TPRI, making it possible to examine predictive validity as well as a second assessment of concurrent validity.

All five tests used to measure WACS validity examine early reading skills and some include subtests similar in name and concept to those given in WACS, providing a stable comparison.

2.3.3 Construct Validity

Construct validity is typically measured with factor analysis or principle component analysis. Because the data was acquired with a computer adaptive test, the large amount of missing data makes a typical factor analysis less useful. Instead, a Rasch Factor Analysis, completed in WINSTEP software, performs a principle component analysis in order to verify that our data is unidimensional. An additional modified factor analysis is run with SPSS for confirmation.

2.4 Reliability

Reliability refers to the consistency of a measure; tests that have adequate reliability will yield more or less the same scores across periods of time and across different examiners. Because WACS is administered on the computer there is no error generated from different examiners. However, error may still be introduced into the resulting final scores through lack of attention to the task at hand, faulty headphones, and disinterest. Because of these concerns, it is important to examine test-retest correlations with a small gap between testing dates as well as the internal consistency of the test.

Computerized adaptive tests differ on measures of test-retest reliability since an individual does not see the exact same test at each time point. The resulting correlation coefficient is regarded as a conservative estimate since content sampling adds an extra degree of error beyond individual performance. CATs also differ on measures of internal consistency. Traditional methods, split-half reliability and Cronbach's Alpha, are statistically inaccurate when applied to a CAT tailored to achievement. Instead, the marginal reliability coefficient provides a better measure of internal consistency by combining measurement error estimated at multiple points on the scale. The resulting coefficient is almost identical to Cronbach's alpha.

3 RESULTS

3.1 Concurrent Validity

Concurrent validity analyses were performed with DIBELS, IRI, ITBS, and TPRI. The Kindergarten WACS combination of tasks includes Blending, Initial Sound, Letter Recognition, Letter Sound, and Vocabulary. The first and second grade combinations include Real Words, Nonwords, Sight Words, Reading Comprehension, and Vocabulary. Overall, correlations between relevant WACS assessments and the associated paper and pencil test are highly significant (Table 2), even with the currently low number of participants taking the ITBS (see Table 1).

Table 1:Number of students completing comparison test.

Assessment	N Kindergarten	N First	N Second
DIBELS	206	142	137
IRI	120	61	126
ITBS		66	69
TPRI	170	155	190

DIBELS Beginning Kindergarten assessment, consisting of Letter Naming Fluency and Initial Sound Fluency, significantly correlates with WACS Kindergarten Skills ($r = .74$, $p < .001$). Similarly, DIBELS Beginning First Grade assessment, consisting of Letter Naming Fluency, Phoneme

Segmentation Fluency, and Nonword Fluency, significantly correlates with WACS 1st grade Skills (r = .72, p < .001) and DIBELS Second grade assessment, consisting of Nonword Fluency and Oral Reading Fluency, correlates with WACS 2nd Grade Skills (r = .61, p < .001).

Patterns or correlations for the IRI are similar to those seen with DIBELS. The IRI includes only one test for Kindergartners, the Letter Naming Fluency test, and this task correlates significantly with WACS Kindergarten Skills (r = .57, p < .001). First grade IRI tasks, Letter Naming Fluency and Nonword Fluency, also correlate highly with WACS 1st grade Skills (r = .74, p < .001). Finally, second graders taking the IRI receive only the RCMB, a reading fluency task. This IRI reading task also significantly correlates with WACS 2nd grade Skills (r = .58, p < .001).

The ITBS includes a number of areas of assessments. For our purposes, comparisons are only made to the Reading subtest for 1st and 2nd graders. ITBS Reading significantly correlates with WACS 1st grade Skills (r = .7, p < .001) as well as with WACS 2nd grade Skills (r = .41, p < .001).

The TPRI is designed for students to receive additional assessments based on previous performance. Only Letter Sound, Blending, Letter Name, and Comprehension are given to all Kindergartners. This combination significantly correlates with the WACS Kindergarten Skills (r = .52, p < .001). The TPRI combination given to all first graders includes Letter Sound, Word Reading, Word Per Minute Rate, and Comprehension Questions. This combination also significantly correlates with the WACS 1st Grade Skills (r = .64, p < .001). Finally, the TPRI combination given to all second graders, including Word Reading, Words Per Minute Rate, and Comprehension, also significantly correlates with WACS 2nd Grade Skills (r = .5, p < .001).

Due to the way the test sequencer works, very few students were given the Segmenting assessment this fall. As a result, Segmenting could not be added to any of the combinations for the grades. However, in order to more thoroughly understand how this task correlates with other reading-related tasks, Segmenting was correlated with two relevant tasks for which the n was over 30. Segmenting correlated significantly with both DIBELS Nonword Fluency (r = .42, p < .05) and IRI Reading (r = .37, p < .05).

Table 2: Relevant correlations between WACS and school-administered assessments.

	WACS Assessments		
	WACS Kindergarten Skills	**WACS 1st Grade Skills**	**WACS 2nd Grade Skills**
DIBELS, Beginning Kindergarten	r = .74, p < .001		
DIBELS, Beginning 1st Grade		r = .72, p < .001	
DIBELS, Beginning 2nd Grade			r = .61, p < .001
IRI, Kindergarten	r = .57, p < .001		
IRI, 1st Grade		r = .74, p < .001	
IRI, 2nd Grade			r = .58, p < .001
TPRI, Kindergarten	r = .52, p < .001		
TPRI, 1st Grade		r = .64, p < .001	
TPRI, 2nd Grade			r = .50, p < .001
ITBS 1st Grade		r = .70, p < .001	
ITBS 2nd Grade			r = .41, p < .001

3.2 Construct Validity

In general, when over 60% of the variance is explained by a single factor, a test is considered to have only one underlying factor. For WACS, 63.5% of the variance is explained by a single factor. Unexplained variance within the 1st factor is 4.5 % (unexplained variance smaller than 5% confirms that there is a single factor). An additional modified factor analysis run with SPSS produces similar results, with the first factor explaining 60.5% of the variance and the next highest factor only explaining 9.7% of the variance. All assessments load strongly on the first factor (all weights above .63) and only Letter Recognition and Letter Sound have weights above .4 on the second factor. In addition, the scree plot indicates a dramatic drop from the first (eigen value of 6.7) to the second factor (eigen value of 1). Finally, another test of the internal coherence of WACS overall is to examine correlations between subtests. Resulting correlations indicate significant relationships among all of the WACS subtests, ranging from r = .38 (between Letter Recognition and Listening Comprehension) to r = .74 (between Letter Sound and Initial Sound), supporting the conclusion that all subtests can be grouped together as a unidimensional test.

3.3 Reliability

Test-retest correlations will be completed in April, 2009, when students take their spring WACS test. The preliminary reliability correlation for WACS Kindergarten Skills, with a sample size of 127, was significant (r = .52, p < .001) as was the reliability correlation for WACS 1st Grade Skills, with a sample size of 85 (r = .73, p < .001).

Internal reliability has already been measured with the marginal reliability coefficient, examining internal test consistency. Reliability for WACS, is very strong (r = .93).

4 CONCLUSIONS

Currently, very few standardized assessments are capable of being used in for pre-K through 2nd grade educational group. Even tests that can be used with these young children often aren't used, likely due to difficulties in keeping young children engaged. WACS has been designed specifically for young children, and by presenting the testing information on the computer, the children are able to stay engaged with the animated characters. Current No Child Left Behind standards require testing beginning in third grade. However, research demonstrates that early detection and intervention are essential for academic success. Identifying struggling students early increases these student's chances of being successful readers and meeting the NCLB requirements. With an assessment that is easy to administer, engaging for the students and provides accurate immediate results, more students are likely to be reading at or above grade level in the future. With validation on the Waterford Assessment of Core Skills completed this coming spring, WACS will become an important part of grade school education.

REFERENCES

Adams, Marilyn, Jager. (1994). Beginning *to Read*: T*hinking and Learning about Print*. Cambridge, Massachusetts: The MIT Press.

Beck, Isabel, L., McKeown, Margaret, G., Kucan, Linda. (2002). *Bringing words to life: robust vocabulary instruction*. New York, New York: The Guilford Press.

Carroll, J. B., Davies, P., and Richman, B. (1971). *The American Heritage Word Frequency Book*. New York: Houghton Mifflin.

Evans, Mary Ann. (2005 June). *Phonological awareness and the acquisition of alphabetic knowledge*. Paper presented at the twelfth annual meeting of Society for the Scientific Study of Reading, Toronto, Canada.

Ganske, Kathy. (2000). *Word journeys*: a*ssessment-guided phonics, spelling, and vocabulary instruction*. New York, New York: The Guilford Press.

Good, R. H., & Kaminski, R. A. (2003). *Dynamic indicators of basic early literacy skills, 6th edition*. Eugene: Institute for the Development of Educational Achievement, University of Oregon

Hoover, H. D., Dunbar, S. B., Frisbie, D. A., Oberley, K. R., Ordman, V. L., Naylor, R. J., Lewis, J. C., Quails, A. L., Mengeling, M. A., & Shannon, G. P. (2003). *The Iowa Tests, guide to research and development, Forms A and B*. Itasca, IL: Riverside Publishing.

Snow, C. E., Burns, S., & Griffin, P. (1998). *Preventing Reading Difficulties in Young Children.* Washington, D.C., National Academy Press.

Software and Information Industry Association. (2008). CODiE Awards, Best Course/ Classroom Management Solution and Best Science Instructional Solution. http://www.siia.net/codies/2008/winners

Stemach, J. & Williams, W. B. (1988). *Word Express: The First 2,500 Words of Spoken English*. High Noon Books.

Wilson, Barbara, A. (2005, November). *Decoding*. Symposium conducted at the meeting of the International Dyslexia Association, Denver, Colorado.

Wilson, Barbara, A. (1996). *Wilson Reading System Instructor Manual*. Boston, Massachusetts: Wilson Language Training Corp.