Meaningful measurement: Is fluency a valid and reliable measure of reading progress for struggling adolescent readers? Examining the efficacy of Pearson's AIMSweb for struggling adolescent readers

Elizabeth A. York

University of New Hampshire, Durham

Follow this and additional works at: https://scholars.unh.edu/thesis

Recommended Citation
https://scholars.unh.edu/thesis/185

This Thesis is brought to you for free and open access by the Student Scholarship at University of New Hampshire Scholars' Repository. It has been accepted for inclusion in Master's Theses and Capstones by an authorized administrator of University of New Hampshire Scholars' Repository. For more information, please contact nicole.hentz@unh.edu.
NOTE TO USERS

This reproduction is the best copy available.

UMI
MEANINGFUL MEASUREMENT: IS FLUENCY A VALID AND RELIABLE MEASURE OF READING PROGRESS FOR STRUGGLING ADOLESCENT READERS?

EXAMINING THE EFFICACY OF PEARSON'S AIMSWEB FOR STRUGGLING ADOLESCENT READERS

BY

ELIZABETH A. YORK

B.A. University of New Hampshire, 1989
M.Ed. University of New Hampshire, 1990

THESIS

Submitted to the University of New Hampshire
in Partial Fulfillment of
the Requirements for the Degree of
Master of Education
in
Reading

September, 2010
This thesis has been examined and approved.

Thesis Director: Ruth Wharton-McDonald, Ph. D.
Associate Professor of Education

Sharon Nodie Oja, Ph.D.
Professor of Education

Maureen Barbieri, M.S.T.
Adjunct Professor
Department of Education

10 August, 2010
Date
TABLE OF CONTENTS

LIST OF FIGURES ......................................................... iv

ABSTRACT ........................................................................ v

CHAPTER PAGE

I INTRODUCTION ............................................................... 1
Defining the Problem .................................................... 2
The Challenge for Teachers ........................................... 4
Monitoring Literacy Growth in High Schools: Pearson’s
AIMSweb ................................................................. 5
Summary of Background Research ............................... 20

II THE RESEARCH STUDY ................................................ 21
Rationale ................................................................. 21
Research Questions .................................................... 21
Method ................................................................. 22
Procedure .............................................................. 23
Data analysis and Results ........................................... 25
Limitations of this Study ............................................. 32

III CONCLUSIONS AND IMPLICATIONS ............................ 33
REFERENCES ............................................................... 38
APPENDIX ................................................................. 42
APPENDIX 1 IRB Approval Letter ................................. 43
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Weekly change in number of words read correctly for Melissa during the 13 week study</td>
<td>28</td>
</tr>
<tr>
<td>2 Weekly change in number of words read correctly for Evan during the 13 week study</td>
<td>28</td>
</tr>
<tr>
<td>3 Weekly performance on R-CBM prompts for the study cohort</td>
<td>29</td>
</tr>
</tbody>
</table>
ABSTRACT

MEANINGFUL MEASUREMENT: IS FLUENCY A VALID AND RELIABLE MEASURE OF READING PROGRESS FOR STRUGGLING ADOLESCENT READERS?

EXAMINING THE EFFICACY OF PEARSON’S AIMSWEB FOR STRUGGLING ADOLESCENT READERS

by

Elizabeth A. York

University of New Hampshire, September, 2010

A growing body of research draws into question the validity and reliability of measurements of fluency as indicators of literacy growth for older, struggling readers (Paris, Carpenter, Paris and Hamilton, 2005; Scammacca, Roberts, Vaughn, Edwards & Torgesen, 2007; Yovanoff, Duesbery, Alonzo & Tindal, 2005). Nevertheless, many high schools utilize such measures. This study examines the validity and reliability of Pearson’s AIMSweb as a tool for assessing the literacy development of 17 9th grade students with special education identifications. Evidence from this study draws into question the reliability and validity of AIMSweb measures for this population. Additional research with a larger population of struggling adolescent readers is warranted.
CHAPTER I

INTRODUCTION

"Reading furnishes the mind only with materials of knowledge. It is thinking that makes what we read ours" (Locke, 1689).

Reading is not a duty, and has consequently no business to be made disagreeable” (Birrell, 1899).

We get no good
By being ungenerous, even to a book,
And calculating profits
--so much help
By so much reading. It is rather when
We gloriously forget ourselves and plunge
Soul-forward, headlong into a book’s profound
Impassioned for its beauty and salt of truth—
“Tis then we get the right good from a book”
(Browning, 1856)

In the passages above, we are reminded of the joy of reading—plunging "headlong into a book’s profound.” Many of us have experienced this rich wonder, and for many of us, it first happened in high school where one particular philosopher, poet or thinker’s words reached out and transformed our view of the world.

For many high school students today, however, these experiences are not happening. For these students, high school is a daily frustration as they are faced with texts that are either well above their reading levels or of very little interest to them (Deshler, 2005). Concerns over the state of adolescent
literacy have increased in the last decade as data from standardized measures such as the National Assessment of Educational Progress (NAEP) indicate widespread skill deficits in adolescent readers. 2009 NAEP testing data indicate that 70% of students entering 9th grade are reading below the proficient level (National Center for Educational Statistics, 2009).

These reading deficits do not disappear in the 9th grade. Biancarosa, Nair, Deshler and Palicsar (2007) suggest that “At-risk high school students across the United States are failing on measures of reading at epidemic rates” (cited in Deshler, 2007, p. 2). Forty percent of students who graduate from high school lack the literacy skills employers seek (Deshler, 2007, p. 2). Furthermore, these struggling readers are entering a work force that will demand that they possess sophisticated reading and thinking skills (Deshler, 2007). Such statistics have garnered the attention of teachers, administrators and parents nationwide, provoking an important discussion regarding literacy instruction and assessment for struggling adolescents.

**Defining the Problem**

Ultimately, every struggling adolescent reader has difficulty with reading comprehension. However, the variety of underlying causes for these reading comprehension difficulties makes helping struggling adolescent readers a difficult proposition. . .The further a student has progressed in the educational system, the more important it is that we intervene in targeted ways (Deshler, 2007 p 35).

In order to intervene appropriately, educators must have access to reliable, valid and meaningful assessments. Most readers who struggle in high school began to struggle as early as the 4th grade. Sadly, the gap between successful and struggling readers increases over time (Stanovich,
1986), leaving those readers who continue to struggle with literacy in high school well behind their peers.

The causes of those struggles are as diverse as the students themselves. For some, it is a lack of practice that leads them to fall behind, for others it is deficits in word knowledge. Some lack awareness of the active reading strategies that enable text comprehension, and others struggle with decoding. The good news is that adolescents do respond to quality instruction; and it is important that they receive quality, targeted instruction in order to succeed (Scammacca, Roberts, Vaughn, Edwards, & Torgesen, 2007, p. 2).

Many researchers suggest specific instructional approaches to help struggling adolescents make gains in literacy. Research has found that given appropriate, targeted instruction in the areas listed below, adolescent readers can make progress (Nair, Deshler, Biancarosa, & Palicsar, 2007; in Deshler, 2007).

- Decoding (for those who still need it)
- Fluency (developed by spending more time reading)
- Vocabulary and background knowledge development
- Direct, explicit comprehension strategy instruction
- Writing instruction
- Information/communication literacy

Specific, targeted instruction is necessary for even slight growth in comprehension to occur (Scammacca et al., 2007, p. 15). In order to make
the most effective instructional choices for struggling readers, teachers need informative, reliable and valid assessments. These assessments play a pivotal role in the design of successful interventions.

**The Challenge for Teachers**

To examine reading assessments and progress monitoring tools in a district is to examine how that district defines literacy. The assessments that district staff choose to use reflect what they value in their students’ learning. Do assessments provide a rich understanding of a student’s literacy needs, or do they simply skim the surface? Are they valid and reliable, or are their data questionable? Poor assessment choices mean misidentification of problems; misidentification of problems means failure to respond appropriately to student needs. “Districts and schools that do not first somehow assess the nature of their students’ literacy strengths and struggles inevitably set themselves up for failure, because without this information, it is impossible to truly match adolescent literacy initiatives to student needs” (Deshler, 2007, p. 35).

The desire for educators to provide struggling adolescent readers with the right interventions based on reliable data must also be contextualized in our current climate of data-driven decision making and high-stakes testing. In our current educational environment assessing and addressing student needs has become a high-stakes, high-cost endeavor. Federal No Child Left Behind legislation requires research-based curricula and assessments. This requirement has driven many districts to purchase commercial products that
provide research-based interventions and assessments that are norm-referenced, promise both validity and reliability in their measures, and are considered quick and easy to use.

The challenge, then, for high school educators is to find assessments that are valid, reliable and meaningful for struggling adolescent readers. If the goal is developing in students the strategies that will lead to successful comprehension, the assessments must provide rich, meaningful data that will inform instruction and monitor student progress in those skill areas necessary for developing comprehension. Furthermore, these measures must be research-based and provide data that can be easily communicated to students, parents and educators.

Adding further complexity to this challenge is the nature of secondary education. It is not uncommon for high school educators to have 130 to 150 students, and few have an aide or assistant in the classroom. Many reading specialists at the secondary level work with large class sizes and do so without assistance. This means that assessment tools must be quick and easy to use. Choices of progress monitoring tools suitable for older readers that are pragmatic scientifically-based and instructionally sound, however, are limited.

**Monitoring Literacy Growth in High Schools: Pearson’s AIMSweb**

Many commercial progress monitoring products today utilize curriculum-based measures of fluency and comprehension to assess literacy growth in adolescents. Most focus on grades K-8. Very few of these tools
address the needs of struggling high school readers. Pearson's AIMSweb is one product that claims to be appropriate for struggling high school readers. This computerized assessment and data management system provides districts with both benchmark and progress monitoring tools and generates data on literacy and mathematics growth for teachers, administrators and superintendents.

The AIMSweb program measures adolescent literacy growth through two measures: weekly R-CBM oral reading passages and bi-weekly or monthly Maze measures. R-CBM passages are reading-curriculum-based measures that are designed to resemble material students might encounter in the classroom. AIMSweb R-CBM measures are one-minute oral fluency probes. Students read a passage at their predetermined reading level aloud for 60 seconds. The teacher scores the reading by drawing a slash through words that are omitted, mispronounced or stated out of order. Students earn a score based upon the number of words they read correctly during those 60 seconds.

The AIMSweb progress monitoring system contains 32 passages (called probes) written at grade levels K-8. Although the passages reach only the 8th grade level, Pearson provides norms tables that extend through the 12th grade. Each set of grade-level probes, from kindergarten through grade 8 has been deemed to be a fair representation of a reading passage one might encounter at the end of that grade-level year. For example, all grade 8 probes are leveled to represent an end-of-grade-8 level of reading difficulty.
AIMsweb also provides Maze measures. These three-minute measures assess student comprehension. In Maze passages, every 7th word in the passage is replaced by a series of three choices that are surrounded by parentheses. The reader must select the word that best completes the sentence she is reading. An excerpt from an AIMSweb Maze probe from the 8th grade level, passage 8p12, illustrates this structure: “Judging by (how, eat, the) looks of the area, Anthony believed (up, he, to) had been told the truth. He (spied, rich, fast) a steep bank through the towering (make, white, days) pines. He commanded his voyageurs, ‘Put (all, bit, the) canoe in there, by that big (find, rock, only)” (Pearson, 2001).

The use of AIMSweb for adolescent struggling readers assumes two things: (a) that measurements of fluency provide valid and meaningful data for older students and (b) the AIMSweb program provides data that are both valid and reliable when tracking the literacy growth of older students. The next section of this paper will consider these two assumptions.

**Assumption One:** Measurements of fluency provide valid and meaningful measurements of reading competency for older students

**Defining Fluency.** Fluency plays a complex role in literacy assessment and instruction. It is identified first in a list of six critical factors underlying proficient reading performance at the late elementary, middle, and high school levels (Torgesen, et. al. 2007). Those factors are:
• Fluency of text reading

• Vocabulary, or the breadth and depth of knowledge about the meaning of words

• Active and flexible use of reading strategies to enhance comprehension

• Background, or prior knowledge related to the content of the text being read

• Higher level reasoning and thinking skills

• Motivation and engagement for understanding and learning from text.

In addition, The National Reading Panel (2000) named fluency as one "of the five pillars of scientific reading instruction" (Allington, 2009, p. 5). Authors such as Richard Allington, Timothy Rasinski, Maryanne Wolf, Sharon Vaughn, Michael Graves, and Connie Juel have acknowledged the importance of this skill to overall reading success. Fluency has several definitions that range from "Reading aloud with accuracy, appropriate speed and expression" (Huey, 1908, p. 140), to "reading accurately while also comprehending what is read" (LaBerge & Samuels, 1974; in Allington, 2009, p. 2) to "reading aloud fast and accurately" (Good & Kaminski, 2002; in Allington, 2009, p. 2).

Richard Allington (2009) defines fluency as "the ability to read in phrases with expression and comprehension" (p. 51). He also explains that, "given the evidence that some children can read accurately and fast while comprehending little, educators must also . . . incorporate measures of comprehension into their assessments of fluency development and their instruction" (p. 5).
The variability in the very definition of fluency poses challenges for teachers. Which definition is most appropriate for measuring growth? Must prosody (reading with expression- and comprehension) be part of that measure, or is reading words quickly and accurately sufficient? John Pikulski and David Chard (2005) make a distinction between surface and deep constructs of fluency that clarifies some of the ambiguity in these definitions. These authors suggest that there are two forms of fluency: surface construct and deep construct. A surface construct of fluency includes the rapid and accurate identification and articulation of words. Readers who have a better surface construct of fluency will read words accurately and efficiently, potentially liberating mental faculties for the job of comprehension.

According to Pikulski and Chard (2005), however, it is not enough to be able to rapidly decode words. Readers need deep construct fluency, which includes both surface fluency as well as a rich understanding of vocabulary (p. 512). Scarborough (2001) develops this idea further by acknowledging that fluency is the coordination of many different skills and understandings. Deep construct fluency engages background knowledge and includes knowledge of print concepts and genre structures. The coordination of these many different skills and understandings will enable fluent reading necessary for understanding. Attention to both surface and deep constructs of fluency enables “efficient, effective word recognition skills that permit a reader to construct the meaning of a text” (Pikulski & Chard, 2005, p. 510).
While surface construct fluency enables a student to be a very accurate word reader, deep construct fluency enables the reader to achieve a rich understanding of what is read. This deep comprehension enables students to contextualize and use the information in a text, which is the goal of most high school literacy endeavors.

How we define fluency affects how we measure it. The differences in the very definition of fluency have resulted in very different approaches to how this skill is measured and taught. Measuring surface construct fluency will provide limited information for a teacher trying to develop comprehension skills in students. For example, Good and Kaminski’s (2002) definition of fluency- reading aloud fast and accurately- has led to their creation of DIBELS (Dynamic Indicators of Basic Early Literacy Skills). This commercial product is used in grades 1 through 3 and measures the number of words a child reads correctly in one minute. Rather than annotating error types, the teacher marks a line through a word read incorrectly or skipped. The student’s score is derived by tallying the number of words read correctly in one minute.

To assess comprehension the teacher asks the student to provide a “retelling” of the passage he has just completed. In a retelling, the student lists the details he remembers from the reading. Each time the child speaks a word that is relevant to the reading passage, the teacher records a point. The total points are translated into a comprehension score.
This retelling keeps the reader's thinking at the very lowest level of Bloom's Taxonomy (a hierarchical model for learning which considers the most basic form of learning to be rote memorization and the higher levels of learning to be those that demand the analysis, synthesis and evaluation of ideas) (Bloom & Krathwohl, 1956). Most high school courses focus their learning at the higher levels of this continuum.

Pressley, Hilden, & Shankland (2005) conducted a study of the adequacy of DIBELS tests. They found that the DIBELS retelling test was not a reliable indicator of comprehension. These authors conclude that, “Based on available data, the fairest conclusion is that DIBELS mis-predicts reading performance on other assessments much of the time and, at best, is a measure of who reads quickly without regard to whether the reader comprehends what is read” (p. 2).

AIMSweb fluency measures reflect a similar demand for surface construct fluency. Retelling is not required in the AIMSweb one-minute readings, so comprehension is not measured. Further, these R-CBM probes do little to inform instruction. During the scoring of the one-minute R-CBM probes teachers are not asked to annotate error types. A separate “Qualitative Features Checklist” is available, if a teacher chooses to use it. This checklist captures whether the student “has an effective strategy for unknown words, reads with expression (attention to prosodic features), self-corrects errors, adjusts pace when complexity or ‘considerateness’ of text changes” and whether “reading errors preserve rather than distort meaning” (Shinn &
Shinn, 2002). No guidance is provided regarding how these observations should be recorded, nor is it clear what value these annotations have for a teacher trying to design instruction for an adolescent reader.

Each criterion listed above will clarify our understanding of the student, but will it do so to the specific degree necessary to design a useful intervention for adolescents? If, for example, a teacher treats the RCBM prompt as a running record and annotates error types, she might find that the student missed most words greater than two syllables or utilized the onset of a word, but guessed at the rime. This information would allow her to design lessons specific to a student’s particular needs with far greater precision than would knowing that the student slowed down when the reading became “less considerate.”

Similarly, the AIMSweb Maze measures ask students to focus on the sentence level of the text. They do not include personal response, analysis, comparison, contrast, or synthesis of the students completing the passages, thus keeping the students’ understanding at the surface level. This surface construct of fluency promotes neither rich understanding nor the utilization of ideas. Nor does it match the sort of reading that high school and the workplace will require of students. Measuring students by this more narrow definition of fluency without an additional comprehension measure is problematic for three reasons.

First, this narrow definition shifts the focus of reading from the creation of meaning to the measurement of rate. Students and teachers see
the final product of the reading experience as a number, rather than an understanding. Second, a narrow definition of fluency that focuses on rate, coupled with the often public aspects of the data gained by these measures, may encourage teachers to narrow the breadth of their literacy instruction to focus on increasing rate rather than providing the rich, meaningful experiences needed to foster adolescent literacy development. The public display of this information may provide subtle pressure for teachers to supplant a more comprehensive approach to reading instruction with one that simply develops rapid reading of text.

This tradeoff can be costly, as many students at the high school level continue to struggle with comprehension. Mary and Anthony Applegate (2009) reported that “assessments of fluency without concurrent assessments of thoughtful comprehension are potentially misleading and damaging” (p. 520). These authors argue that we must consider fluency as a constituent of comprehension, not as a sufficient indication that comprehension has occurred. “What may ultimately be even more detrimental,” they argue, “is the establishment of programs of instruction that divorce fluency and word recognition from comprehension” (p. 520).

Finally, as elaborated below, there is a growing body of evidence that fluency is not a reliable measure of adolescent literacy progress.

**Fluency as a Measure of Adolescent Literacy Progress**

Measurements of fluency, even those at the surface level, have been shown to have a strong correlation with comprehension for students at the
elementary levels (Torgesen, 2006). As it is easy to measure, and appears to serve as a valid proxy measure for comprehension at earlier grades, many districts have selected progress monitoring programs that utilize fluency to monitor the literacy progress of their students. Some of these districts are also utilizing these measures with students in grades 8-10. AIMSweb markets itself as appropriate for any age/grade, “Currently, AIMSweb measures are available for Benchmarking (Universal Screening) K-8 and Progress Monitoring any age/grade. However, many high schools are using AIMSweb materials for intensive progress monitoring at-risk students” (http://www.aimsweb.com).

While fluency is correlated with comprehension in the early grades, (Allington, 2009, p. 6) that correlation weakens in later grades (Yovanoff, Duesbery, Alonzo, & Tindal, 2005). Indeed researchers are finding that fluency is not a sufficient proxy for comprehension for older students “due to the multiple shared processes that may account for relations between fluency and comprehension, such as vocabulary, syntactic knowledge, and background knowledge” (Paris, Carpenter, Paris, & Hamilton, 2005, p. 312). Further, the importance of this skill relative to other literacy skills that students are developing diminishes as students age. For older readers, fluency instruction has not been found to have a significant impact on comprehension (Scammacca, 2007).

A meta-analysis of research studies titled Reading Interventions for Adolescent Struggling Readers suggested effects sizes of only .26 for fluency instruction on adolescent comprehension. “The effects of fluency
interventions on standardized measures and on all measures of reading comprehension were not reliably different from zero” (Scammacca, et al., 2007, p. 16). “For older students the role of fluency instruction generally, and the relative effects of differing instructional approaches for improving reading outcomes, needs additional research” (Torgesen, et al., 2008 p. 65). Measures of fluency for high school readers may be much less valid, and therefore much less informative for teachers than such measures are at the elementary level.

If our assessments guide our instruction, assessments with a narrow focus on surface fluency may have detrimental effects for struggling adolescent readers who will be asked to utilize the information they read rather than simply decode it. “By fourth grade, when reading disability has been formally diagnosed for many students with learning disabilities, so many components of reading are lagging seriously that a focus on fluency alone may be counterproductive” (O’Connor, 2007, p. 33).

Current research suggests that fluency measures do not provide the rich, meaningful data that teachers need to provide targeted instruction for teens. This is an important consideration for teachers and administrators as they select progress monitoring tools for their districts. A second important consideration is the selection of tools that are both valid and reliable.
**Assumption Two:** Pearson's AIMSweb is a valid and reliable measure of literacy progress for high school students

**Questionable Claims.** In the last section, we examined whether fluency is an appropriate measure of literacy growth for older students. Current research has drawn that assumption into question. The second assumption that use of the AIMSweb program implies is that AIMSweb provides valid and reliable assessment instruments. AIMSweb markets itself as a research-based product. There is little evidence, however, supporting the reliability and validity of AIMSweb as a progress monitoring measure for adolescent readers. The AIMSweb website, www.aimsweb.com, provides a collection of research to support the legitimacy of their product. Most of the articles included in that collection, however, are not peer-reviewed studies of the program itself. They are reviews of the efficacy of R-CMB measures in general.

For example, the first research article listed on the AIMSweb site is an executive summary put forth by the Institute for the Development of Educational Achievement titled *Executive Summary of Final Report on Reading First Reading Assessment Analysis: Analysis of Reading Assessment Instruments for K-3*. This report supports the use of R-CBM measure in grades K-3 and provides a list of progress monitoring programs that the Institute found reliable. Surprisingly, AIMSweb is not listed among those recommend programs; their competitor, DIBELS, however, is (Gamse, Jacob, Horst, Bouley, & Unlu, 2008, p. 1).
A second title listed as research supporting the AIMSweb program is titled, "Variables that Affect the Correlation between Fluency and Accuracy with a Measure of Comprehension," written by Michelle K. Hosp, of Vanderbilt University. Beyond its listing on the AIMSweb site, no other publication location is mentioned, making this author's independence from the Pearson company questionable. A third article, "Summary of Reliability Studies for General Outcome Measures of Reading" directs the reader to the AIMSweb Training Workbook (Shinn & Shinn, 2002) which provides a chart that lists a series of studies of the validity of R-CBM measures in general, including no specific examinations of AIMSweb itself. A final article, with the promising title "AIMSweb CBM Tools Meet Scientific Standards for Use in Frequent Progress Monitoring" provides a link to the National Center on Response to Intervention Screening Tools Chart.

The National Center on Response to Intervention (NCRTI) is an organization that is supported by the American Institutes for Research and by researchers from both the University of Kansas and Vanderbilt University. NCRTI is funded through the department of Education's Office of Special Education Programs. The Center's mission is to provide both technical and professional support to districts implementing RTI (Response to Intervention) practices.

The NCRTI provides a variety of resources, including a chart that rates the reliability and validity of RTI programs. In an AIMSweb press release, Pearson claims that its AIMSweb program has received The National Center
for Response to Intervention's (NCRTI) highest rating for commercially available RTI tools; and indeed it has, for some samples of R-CBM measures for four of the nine grade-levels they measure. RCBM measures in only grades 1-3 are considered valid and reliable. Yet, an AIMSweb press release claims that this rating reinforces “the program’s effectiveness, quality and usability. This independent rating from the leading RTI standard-setting organization is outside confirmation that AIMSweb assessments are valid and reliable” (http://www.aimsweb.com/news). No mention is made that this validity and reliability extends no further than the third grade.

Further exploration of the NCRTI sites shows that this organization gave AIMSweb only a “Partially convincing” rating on its sensitivity to student growth. The NCRTI also found that AIMSweb’s predictive validity for its slope of improvement falls between .23 and .43 for grades one through three” (rti4success.org/chart).

AIMSweb also claims high ratings from an organization called Student Progress.org. This site directs the visitor back to the chart available on the NCRTI site. There is no other mention of AIMSweb’s reliability or validity on that site. The NCRTI site and research conducted by Theodore Christ and Scott Christ (2009) elaborated on later in this paper, are the only independent assessments of AIMSweb’s reliability and validity that this researcher could locate.
The Validity and Reliability of AIMSweb Data for Struggling Adolescent Readers

The correlation between fluency and comprehension weakens as students age (Yovanoff et al., 2005). Interestingly, data produced from utilizing fluency measurements with teens tend to exhibit particular characteristics. A first characteristic is score deviation. Many students being monitored using fluency measures exhibit high score deviations from one week to the next. These scores are derived from prompts considered to be of equal difficulty by their publisher. This brings to light the second characteristic in the data, that student scores tended to rise and fall in concert, perhaps indicating that passages are not accurately leveled. These two concerns draw into question the reliability and validity of AIMSweb measures.

Independent research raises some concern regarding the level of error that is presented in R-CBM measures such as DIBLES and AIMSweb:

There are relatively few studies that evaluate the quality of progress monitoring estimates derived from curriculum-based measurement of reading. Those studies that are published provide initial evidence for relatively large magnitudes of standard error relative to the expected magnitude of weekly growth. A major contributor to the observed magnitudes of standard error is the inconsistency of passage difficulty within progress monitoring passage sets (Ardoin & Christ, 2009, p. 266).

The excessive fluctuations in scores and lack of proven reliability in R-CBM measures, according to Ardoin and Christ (2009) can lead to errors in identifying student needs, measuring student progress and planning accurate and useful interventions. The authors argue that R-CBM measures, which are
currently being used to measure the effectiveness of classroom interventions, to set IEP goals and to make important placement and curricular decisions should receive the same scrutiny regarding reliability and validity as general standardized measures such as the Woodcock-Johnson or WISC tests (Ardoin & Christ, 2009, p. 281).

**Summary of Background Research**

Given the challenging and complex nature of adolescent literacy, it is vital that assessments used to monitor literacy growth for these students are valid, reliable and meaningful. Many schools are turning to R-CBM fluency measures such as those provided by the AIMSweb program to monitor adolescent literacy progress. Those measures should be considered with significant skepticism for three reasons. First, while fluency provides valid and reliable measures of literacy achievement for elementary students, the effectiveness of this measure as a tool for progress monitoring diminishes as students age. Second, there is no empirical evidence that the AIMSweb measures of reading growth are valid and reliable beyond the third grade level. Third, the standard error for AIMSweb measures for individual students is very large. These three concepts formed the backdrop for this study. This study examined the efficacy of Pearson=’s AIMSweb as a progress monitoring tool for a group of struggling adolescent readers.
CHAPTER II

THE RESEARCH STUDY

Rationale

Struggling adolescent readers pose a special challenge to educators. They require instruction targeted to their specific literacy needs. In order to target instruction effectively, teachers must rely upon valid, reliable and meaningful assessments. Many high schools have turned to Pearson’s AIMSweb as a vehicle for monitoring the literacy growth of their students. This study was designed to consider the reliability and usefulness of this program as a progress monitoring tool for struggling ninth grade readers who have a special education designation.

Research Questions

Is Pearson’s AIMSweb a valid, reliable and useful measure of literacy growth for struggling adolescent readers? This study will evaluate AIMSweb data for 17 9th grade students with the following suppositions:

- If the leveling of AIMSweb R-CBM passages is consistent and reliable, large gains or losses in fluency ratings should not occur A) in individual students from one week to the next, and B) simultaneously for entire groups of students reading the same passage on the same day.
If the AIMSweb Maze measures are valid measures of reading comprehension, student scores on other valid, standardized comprehension measures such as the Gates-MacGinitie Test of Reading Comprehension should correlate positively with scores on the AIMSweb Maze comprehension measures.

Method

Participants

Seventeen 9th grade special education students participated in this study. These students were enrolled in both a mainstream, heterogeneously-grouped Language Arts class that met for 223 minutes weekly and a supplemental reading course that provided an additional 223 minutes of direct reading instruction each week. These students were placed in this supplemental reading class based upon their eighth grade scores on the NECAP (The New England Common Assessment Program) reading test and the NWEA (North West Evaluation Association) MAP (Measure of Academic Progress) reading test.

This study took place in a small northern New England city over the course of 13 weeks between November, 2009 and early May, 2010. Data from the 2000 census indicate that the city has a median household income of $42,447; 9.3% of those under the age of 18 live below the poverty line. The high school where this study took place houses 2000 students, 88% of whom are white. This district, which serves a community of 42,255 residents had recently been labeled a District in Need of Improvement. As part of its plan
to increase student achievement, administrators have purchased Pearson’s AIMSweb to enable the tracking of student progress K-10. AIMSweb costs approximately $5.00 per student. This district is currently monitoring the progress of 900 students in grades K-10.

**Procedure**

At the high school level, the data derived from AIMSweb are used during Literacy Team meetings. During these meetings, staff members discuss the progress of individual students and make changes to existing interventions. These changes may include adjustments in course content such as reading materials or instructional approaches, or broader changes such as an adjustment in a student’s daily schedule. Some teachers are utilizing AIMSweb data to set progress goals on IEPs, (Individualized Education Programs) making these data part of a legal document to which teachers and specialists must adhere.

Students participating in the study were given weekly AIMSweb R-CBM and Maze reading assessments over the course of 13 weeks. These readings were part of their weekly classroom routine in their special education reading course and reflected practices that were familiar to them. Results were tracked through AIMSweb’s progress monitoring system, an online data warehouse that enables specialists and administrators to examine individual and group data.

Later in the spring, as part of a building-wide initiative that encompassed all 9th and 10th graders, students were given the Gates
MacGinitie Reading Test, Level 7/9, form S. The data from this test were utilized in this study.

The tracking of AIMSweb scores for this study began in November, two months after the system was introduced to school district personnel who provide interventions for struggling readers K-12. This gap gave both the teacher and the students ample time to become familiar with the program protocols before data were collected for this study. For the purposes of this study, “the teacher” refers to the author of this paper.

The teacher received several training sessions prior to utilizing the AIMSweb program. One two-hour session was devoted to orienting staff to the program to ensure fidelity. All staff received practice scoring prompts to develop consistency in test administration. A second, two-hour session was devoted to utilizing the AIMSweb Progress Monitoring System, the computerized database for storing and tracking student scores. There were also two district-wide follow-up meetings for troubleshooting, as well as one optional refresher training session for staff.

During this study, the students were assessed individually using the AIMSweb R-CBM one minute passages once each week. They were often asked to draft or state a brief retelling of the passage after the reading was completed. Students were also given the three-minute Maze passages once per week. Following these passages, students were often asked to complete a brief retelling as well.
Students were administered the Gates-MacGinitie Reading Test during the first week of April, 2010. These results were compared with their AIMSweb Maze comprehension measures during that same time period.

Data Analysis and Results

Data Analysis

Several questions were considered in this study:

• Does mean growth in number of words read correctly for this group of students correlate with the expected mean provided by AIMSweb?

• How great is the standard deviation in the number of words read correctly from one prompt to the next for these students?

• What is the range in measures of weekly gains or losses in numbers of words read correctly?

• Is there a correlation between fluency and comprehension measures provided by AIMSweb?

• Is there evidence that students’ gains and losses are dependent on the particular prompt they are reading—drawing into question the leveling of the text?

• What is the correlation between AIMSweb comprehension scores and comprehension scores from the Gates-MacGinitie Reading Test, a second nationally-normed, standardized measure of comprehension?

In order to preserve consistency in the data, only those students who were reading prompts at the 8th grade level were included in this data
analysis. This modifies the number of students participating from 17 to 13. All statistics were generated using Fathom Dynamic Data software.

**Results**

The data gathered from this sample indicated that the AIMSweb progress monitoring system was not a reliable indicator of growth for these 13 9th grade students. The findings are illustrated below.

Does mean growth for this group of students correlate with the expected mean provided by AIMSweb? A first data point considered was the mean rate of growth for the sample students in relation to the rate of growth predicted by Pearson. Pearson predicts a gain of .4 words read correctly each week for a student reading at the 8th grade level. This score is based on the mean growth score of the entire norming group for that grade level. Essentially the mean score reflects the mean of all the weekly growth and loss scores for students in the norming group. Comparing individual growth per week against these mean rates of growth, however, reveals high levels of standard error in the data (Ardoin & Christ, 2009). Ardoin and Christ argue that positive mean growth data for a group often mask alarmingly large deviations in the scores of individual students from prompt to prompt, suggesting issues with the leveling of the material. Further, the mean group score can mask more important data trends such as gains and losses over time. A large group with a positive mean score could have actually lost ground over time, as did the student cohort considered for this study. Their mean score did not reflect that loss.
A t test of the sample students’ average weekly scores showed a mean change in student scores of +.67 correct words per week. The sample scores show a strong correlation with the expected change in scores of +.4 words per week ($p=.89$). Initial inspection of this data seems promising. This mean score correlates well with the expectations of the standardized AIMSweb measure and shows greater growth than the program expected. These data would seem to suggest that students are thriving. This mean score, however, can be misleading for two reasons, a) it doesn’t capture growth over time and b) it doesn’t take into account the growth (or lack thereof) in individual student scores—which is the heart of progress monitoring. Student samples that follow will illustrate these points.

How great is the deviation in the number of words read correctly from one prompt to the next for these students? What is the range in measures of weekly gains or losses in numbers of words read correctly?

Over the course of the 13 week study, R-CBM fluency measures of individual reading progress varied widely. The range of these scores was +1 - 40 cwpm with standard deviations of 19.2 words per week. Figures 1 and 2 on the following page represent typical students’ scores over the course of the thirteen week study.
Figure 1. Weekly change in number of words read correctly for Melissa over the 13 week study.

Figure 2. Weekly change in number of words read correctly for Evan over the 13 week study.
Evan's and Melissa's charts exhibit the same wide variation in scores that most students in this study exhibited. The wide variations of +/- 19 words per week on average suggest issues with either the leveling of the prompts or the validity of fluency as a measure of literacy growth for these students. If the prompts were equally difficult, such wide variations in weekly scores should not occur. Further, with such wide variations in data, teachers will find it difficult to trust these data as a basis from which to judge the efficacy of an intervention.

Is there evidence that students' gains and losses are dependent on the particular prompt they are reading? Further examination of group data reveals that student scores appear to rise and fall in concert on various prompts. Figure 3 below illustrates how students rose or fell on particular prompts.

Figure 3: Weekly performance on R-CBM prompts for the study cohort.
To test whether the apparent rising and falling of scores was dependent on the passage itself, a chi-square analysis was run. This form of analysis requires a significant amount of data. As the data sample in this study is small, the findings of this calculation cannot be transferred to a broader population.

Though the data set is small, it suggests that a relationship between each prompt and the collective rising and falling of student scores is likely. The p value of .00067 for the likelihood of the measures being independent of each other is very weak, suggesting that variations in the prompts themselves may be influencing students' scores. This is further confirmed by a study of the readability levels of many of the passages. Both Chall and Kincaid readability calculators were used to examine the level of difficulty presented by individual AIMSweb R-CBM passages. These calculators provide an assessment of the level of difficulty a passage represents. The score is conveyed as a grade level. All AIMSweb passages tested with these calculators should represent a passage written with an end of 8th grade level of difficulty. Only 1 of the 8 passages examined represented this level of difficulty. The most common difficulty level was 7.5, and the range of grade-level scores for all passages varied from early 4th grade to the late 8th
grade. This suggests that further study of the quality of the leveling of these passages is merited.

*Is there a correlation between fluency and comprehension measures provided by AIMSweb?* Another important component of the AIMSweb program is the use of the Maze prompts to measure comprehension. These prompts are optional at the New England school sampled, but were utilized in this study. The Maze passages were administered weekly to the students in the sample. Over the course of the study, student fluency rates and comprehension rates varied widely. During the week of March 13th, 2010, for example, 13 students were near or above target in their comprehension scores, while 3 were below. In contrast, that same week only 3 students were near or above their target for fluency with 13 falling below.

These data illustrate the complicated nature of monitoring literacy progress for teens. Using fluency to measure the literacy progress of teens assumes that there is a positive correlation between fluency and comprehension; here these measures contradict. This trend continued to varying degrees throughout the 13 week study supporting the arguments of Scammacca, et al. (2007), Paris, et al. (2005), and Yovanoff, et al. (2005) that the correlation between fluency and comprehension declines as students age making it an unreliable measure of adolescent literacy growth.

*Do AIMSweb comprehension scores correlate well with the Gates-MacGinitie test, a second nationally-normed, standardized measures of comprehension?* To further investigate the accuracy of the AIMSweb
comprehension readings, student percentile scores from this program were compared with their percentile scores from the Gates-MacGinitie Reading Test Level 7/9. Given that both tests measure comprehension, there should be some relationship between the two scores sets. As AIMSweb provides very limited choices for percentile ranges—10, 25, 50, and 75—the ability to make specific correlations is limited, so a 1:1 relationship cannot be expected. There should, however, be some correlation between the two measures of comprehension.

The test yielded an $r^2$ of .17 suggesting that only 17% of the variation in the Gates Scores can be explained by performance on the AIMSweb measure. As the sample size is small, one cannot assume that this is true for broader populations; nonetheless, this weak correlation suggests that further investigation of this relationship between these two measurement devices may be merited.

**Limitations of this Study**

There are two primary limitations to this study. The first limitation is the sample size. Due to the small number of students involved in the study, the data cannot be considered applicable to broader populations. Second, attendance was a significant issue for three of the students who participated in the study. Frequent absences impacted the number of data points that could be gathered for those students. Despite these limitations, the data gathered from this study provoke questions regarding the validity and reliability of AIMSweb that merit further research.
Due to the prevalence of commercial progress monitoring programs today, it is important to consider their efficacy. Do these products serve districts well? Do their findings inform instruction in a meaningful manner? Are they fair and reliable measures of the complex process of meaning making for adolescents? This study raises some questions regarding the efficacy of AIMSweb fluency measures for special education students at the secondary level. These data suggest that further inquiry into the usefulness of AIMSweb measures for struggling high school students is needed.

The first and most broad consideration of AIMSweb should begin with the examination of the definition of fluency it reflects. Surface-level fluency involves reading words quickly and accurately. A Deep Construct model of fluency demands rich understanding of vocabulary and text structure in order to develop an understanding of a text which can be applied to a variety of tasks. The AIMSweb program's R-CBM measures focus attention on surface-level fluency which does not represent the intellectual demands that will face high school students in the classroom.

A second concern is the phenomenon of student scores rising and falling in concert. Data suggest that there may be some variability in the difficulty level of the prompts, whether in readability or content, and this
may influence individual student scores. During a meeting with three reading specialists who are using the AIMSweb prompts in the selected school, the concern over the uniform rising and falling of scores was voiced by all, suggesting that this phenomenon extends beyond the small special education population included in this study. Whether these prompts are leveled reliably and are, therefore, valid measures of student progress, will be an important aspect for districts who are considering purchasing the AIMSweb program to evaluate.

A third area that merits further inquiry is whether the passages are valid and reliable measures of student comprehension. Few independent research studies have been conducted to test the validity and reliability of this product. Given the popularity and cost of this product, this research is needed. R-CBM fluency and comprehension measures are of questionable validity and reliability for older readers. As students age, the correlation between fluency and comprehension weakens significantly, making the relevance of these data for designing interventions questionable.

A final concern is that these data are also public. This, coupled with the high stakes nature of No Child Left Behind legislation, could lead to teachers narrowing the focus of their instruction in order to increase student rate of reading, a skill that shows little relation to comprehension for readers at the secondary level.

The monitoring of student literacy development is an important aspect of effective teaching. The era of high stakes testing and federal mandates
adds even greater importance and power to this teaching practice. The challenge for high school teachers of special education students is finding accurate, reliable, research-based measures of progress. As data from progress monitoring programs often lead teachers to modify instruction and to make important placement decisions for their students, the validity and reliability of progress monitoring systems must be tantamount in the minds of those establishing progress monitoring protocols in districts. Commercial producers of progress monitoring systems have an absolute obligation to hold themselves to the highest empirical standards for validity and reliability, knowing the power that rests within these measures.

AIMSweb is considered very effective and reliable for elementary students, especially those in grade three. Its efficacy for high school special education students, whether due to the decreasing correlation between comprehension and fluency as students age, or due to structural issues within the program itself, is not yet proven. This should be carefully considered prior to purchasing this expensive program.

“Older, struggling readers are extremely complex. . .to meet their needs we need to take a closer, more sophisticated look at their strengths, needs and preferences. . .we have to see them engaging in literate tasks in a variety of contexts, . . .and for a variety of purposes” (Fisher & Ivey 2006). High schools that choose to use the AIMSweb program should be aware of its limitations and should include it, as did the high school utilized for this
study, as only one of many data points when considering the complex needs of their students.

This evidence does not suggest that fluency is not important. It does, however, clarify that for older, struggling readers, fluency assessment and instruction should be part of a broad approach that includes frequent reading experiences at appropriate levels of difficulty and assessments that involve demonstration of understanding of a text, echoing Pikulski and Chard's Deep Construct fluency model (2005). Districts planning to purchase programs for monitoring the literacy growth of their high school students should take steps to ensure that fluency readings are coupled with comprehension assessments and are part of curricula that includes rich experiences with text, explicit teaching of reading strategies, vocabulary development and writing (NH Literacy Action Plan, 2007).

It is important that students with LD or reading difficulties receive appropriate intervention. The difficulty of the task should not be underestimated, and effective instruction is only one piece of the larger puzzle, albeit an important piece. . .Older students with reading difficulties can benefit from well-designed, effectively delivered intervention (Scamacca et al., 2007, p. 15).

If our assessments mirror our goals, districts that utilize AIMSweb measures are seeking to create students with a strong memory for detail and accurate surface-level understanding. The stakes are high for struggling adolescent readers, and teachers require meaningful, reliable tools in order to provide the interventions that will enable these students to succeed. Most high school and college classes will ask students to use the information they read for a particular purpose. How will data such as the number of words
read correctly in a minute or the number of words chosen correctly in a Maze prompt inform instruction? Given the complex literacy needs of struggling adolescent readers and the lack of empirical evidence supporting AIMSweb's efficacy for teens, one must question whether the investment of time and money required to manage the AIMSweb system in high schools is worthwhile.

If our goal is to create readers who can read for meaning and can utilize the knowledge that they gain to enable their own creative thoughts and actions, then we must use assessments that measure the skills students require to achieve these goals. High school students are often asked to explore a text, to think deeply about it, and respond meaningfully to it—to essentially reach into a book's profound and discover its "beauty and salt of truth." We must ask ourselves whether measuring the number of words read correctly in one minute is the best path to get them there.
REFERENCES


APPENDIX 1

IRB APPROVAL LETTER

University of New Hampshire

Research Integrity Services, Office of Sponsored Research
Service Building, 51 College Road, Durham, NH 03824-3585
Fax: 603-862-3564

15-Feb-2010

York, Elizabeth
Education, Morrill Hall
92 Carter Hill Road
Concord, NH 03303

IRB #: 4703
Study: Meaningful Measurement: The Efficacy of AIMSweb and Read Naturally in Measuring Comprehension in Struggling Adolescent Readers
Approval Date: 15-Feb-2010

The Institutional Review Board for the Protection of Human Subjects in Research (IRB) has reviewed and approved the protocol for your study as Expedited as described in Title 45, Code of Federal Regulations (CFR), Part 46, Subsection 110.

Approval is granted to conduct your study as described in your protocol for one year from the approval date above. At the end of the approval period, you will be asked to submit a report with regard to the involvement of human subjects in this study. If your study is still active, you may request an extension of IRB approval.

Researchers who conduct studies involving human subjects have responsibilities as outlined in the attached document, Responsibilities of Directors of Research Studies Involving Human Subjects. (This document is also available at http://www.unh.edu/osr/compliance/irb.html.) Please read this document carefully before commencing your work involving human subjects.

If you have questions or concerns about your study or this approval, please feel free to contact me at 603-862-2003 or Julie.simpson@unh.edu. Please refer to the IRB # above in all correspondence related to this study. The IRB wishes you success with your research.

For the IRB,

Julie F. Simpson
Manager

cc: File
Cioffi, Grant