Examining the Validity of the CUBED Narrative Language Measures Listening and Reading Subtests

Trevor David Read
*Brigham Young University*

Follow this and additional works at: [https://scholarsarchive.byu.edu/etd](https://scholarsarchive.byu.edu/etd)

*BYU ScholarsArchive Citation*

Read, Trevor David, "Examining the Validity of the CUBED Narrative Language Measures Listening and Reading Subtests" (2023). *Theses and Dissertations*. 9962.
[https://scholarsarchive.byu.edu/etd/9962](https://scholarsarchive.byu.edu/etd/9962)

This Thesis is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.
Examining the Validity of the CUBED Narrative Language Measures

Listening and Reading Subtests

Trevor David Read

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Connie Summers, Chair
Shawn Nissen
Douglas Petersen

Department of Communication Disorders
Brigham Young University

Copyright © 2023 Trevor David Read
All Rights Reserved
ABSTRACT

Examining the Validity of the CUBED Narrative Language Measures
Listening and Reading Subtests

Trevor David Read
Department of Communication Disorders, BYU
Master of Science

Purpose: The purpose of this study was to examine the concurrent and predictive validity of the CUBED Narrative Language Measures Listening and Reading subtests to evaluate their psychometric properties as defined by Deno (2003). Method: Evidence of concurrent validity for the NLM Listening and Reading subtests was examined using data from 1,146 preschool through third grade students. The NLM subtests were compared to similar subtests from other assessments administered at approximately the same time. Evidence of predictive validity was examined using data from these same NLM subtests from 1,512 kindergarten through third grade students, including 71 third grade students for the examination of sensitivity and specificity. This was accomplished by examining how the students’ NLM assessment scores related to their future performance on tests measuring similar constructs. Sensitivity and specificity were determined by examining how the well the NLM subtests identified students performing at or below grade level according to state proficiency tests. Measures used to determine concurrent and predictive validity included the following: curriculum-based assessment for writing, narrative language sample, expository language, Clinical Evaluation of Language Fundamentals 2 (CELF-P 2), Renfrew Bus Story, Measures of Academic Progress (MAP), and Wyoming Proficiency subtests for Wyoming Students (PAWS). Results: Results indicate that the NLM Listening and NLM Reading had moderate to strong correlations with all criterion measures for concurrent validity. Results also indicated the NLM Listening and NLM reading, separately and combined, accounted for meaningful variance in criterion measures used for predictive validity. Combined, the NLM Listening and Reading had fair to excellent sensitivity and specificity for criterion measures, with 77% sensitivity and 100% specificity for end-of-year state reading assessment performance. This study provides evidence that the NLM Listening and Reading can accurately measure and predict oral language and reading outcomes in students preschool through third grade, indicating it as a useful curriculum-measure as defined by Deno (2003). Future research should examine the extent to which the entire CUBED assessment, including the decoding and dynamic assessment subtests, have concurrent and predictive evidence of validity.

Keywords: reading comprehension, oral language, curriculum-based measures
ACKNOWLEDGMENTS

I would like to thank Dr. Douglas Petersen for guiding me in writing this thesis as my thesis chair while he was working for BYU and as a member of my committee when began working at the University of Wyoming. I would like to thank Dr. Connie Summers for her guidance as she stepped in as my thesis chair in place of Dr. Petersen. I would also like to thank Dr. Shawn Nissen for being a part of my committee and assisting me in the process of formatting in preparation for defense.
TABLE OF CONTENTS

TITLE PAGE........................................................................................................................................... i
ABSTRACT.................................................................................................................................................. ii
ACKNOWLEDGMENTS ............................................................................................................................. iii
TABLE OF CONTENTS............................................................................................................................... iv
LIST OF TABLES........................................................................................................................................ vi
DESCRIPTION OF THESIS STRUCTURE AND CONTENT ................................................................. vii

Introduction............................................................................................................................................... 1

Curriculum-Based Measures..................................................................................................................... 2

Validity of Curriculum-Based Measures ................................................................................................ 2

Construct Validity....................................................................................................................................... 3

The CUBED Narrative Language Measures Listening and Reading Subtests.................................... 6

Statement of the Problem ........................................................................................................................ 10

Statement of the Purpose ......................................................................................................................... 11

Research Questions ................................................................................................................................. 11

Method...................................................................................................................................................... 11

Participants............................................................................................................................................... 11

Procedures............................................................................................................................................... 12

Concurrent Validity ................................................................................................................................ 12

Predictive Validity................................................................................................................................... 12

Measures .................................................................................................................................................. 13

Curriculum-Based Assessment for Writing ............................................................................................ 13

Narrative Language Sample (Frog, Where Are You?) ......................................................................... 14
Expository Language ........................................................................................................ 14
Clinical Evaluation of Language Fundamentals Preschool 2 (CELF-P 2) ................. 15
Renfrew Bus Story ........................................................................................................... 15
Measures of Academic Progress (MAP) ......................................................................... 15
Wyoming Proficiency Assessments for Wyoming Students (PAWS) ......................... 16
Results .............................................................................................................................. 17
Concurrent Validity .......................................................................................................... 17
Predictive Validity ............................................................................................................ 18
Discussion ....................................................................................................................... 18
Concurrent Validity .......................................................................................................... 19
Predictive Validity ............................................................................................................ 22
Validity of the Narrative Language Measures ............................................................... 23
Clinical Implications ....................................................................................................... 25
Limitations and Future Research .................................................................................... 25
References ....................................................................................................................... 26
Tables ............................................................................................................................... 32
APPENDIX A: Annotated Bibliography ........................................................................... 41
APPENDIX B: Consent/Institutional Review Board Approval Letter ......................... 53
LIST OF TABLES

Table 1  Unattenuated/Corrected(Uncorrected) Correlation Coefficients Between the
Fall NLM Listening and Language-Related Criterion, Measures Evidencing
Concurrent Validity ........................................................................................................32

Table 2  Unattenuated/Corrected(Uncorrected) Correlation Coefficients Between the Fall
NLM Reading and Language-Related Criterion Measures, Evidencing Concurrent
Validity ..........................................................................................................................36

Table 3  Regression Analyses Using Fall NLM Listening Raw Scores to Predict Winter
MAP Performance Evidencing Predictive Validity .........................................................37

Table 4  Regression Analyses Using Fall NLM Reading Raw Scores to Predict Winter MAP
Performance Evidencing Predictive Validity ....................................................................38

Table 5  Corrected (Uncorrected) Correlations Between the Fall NLM Listening and
Reading Scaled Score Composite and Winter MAP .......................................................39

Table 6  Sensitivity and Specificity of NLM Listening and NLM Reading Combined Third
Grade Fall Assessment for End of Year PAWS Reading Assessment and Middle of
Year MAP Reading Assessment ....................................................................................40
DESCRIPTION OF THESIS STRUCTURE AND CONTENT

This thesis, *Examining the Validity of the CUBED Narrative Language Measures Listening and Reading Subtests*, was written to adhere to traditional thesis requirements as outlined by the Brigham Young University McKay School of Education. Appendix A consists of the annotated bibliography. An International Review Board (IRB) exemption approval letter provided by the University of Wyoming pertaining to the data used within this thesis is included in Appendix B.
Introduction

The simple view of reading proposed by Gough and Tumner (1986) postulates that reading comprehension is the product of decoding and language comprehension. Decoding refers to one’s ability to recognize words with automaticity and with correct letter-sound correspondence, and language comprehension refers to the ability to understand the words being read. For example, a person might be able to decode the words of a foreign language, recognizing the orthographic representations of that language’s sounds, but fail to understand the content of what is being decoded. Conversely, a person might have strong oral language skills in a foreign language but fail to decode that language’s written form. If either decoding or language comprehension difficulties are present, reading comprehension is negatively impacted. Many researchers provide data that support this indication (Catts, 2018; Catts et al., 2005; de Jong & van der Leij, 2002; Lonigan et al., 2018). For example, Catts et al. (2006) reported that all variance in young students’ reading comprehension was the result of poor decoding abilities, poor language comprehension, or both. Therefore, the simple view of reading accurately describes the importance of considering both decoding and language comprehension in any effort to understand why a student may have reading comprehension difficulties (Hogan et al., 2014).

The National Assessment of Educational Progress (NAEP, 2022) found that 63-69% of fourth and eighth grade students read below a proficient level. In the context of the simple view of reading this NAEP data suggests that we need to identify whether these students are struggling with decoding and/or language comprehension and to then do something about it. The earlier we can identify these students who have decoding and/or language difficulty and begin monitoring their progress, the better. It is necessary to provide systematic and focused reading intervention early to assist the student overall in their education (Lyon, 2003). To do this we need accurate
and valid universal screeners and progress monitoring tools that measure decoding and language to identify reading difficulties (Gilbert et al., 2012; Petersen & Stoddard, 2018). These universal screeners and progress monitoring tools are often referred to as curriculum-based measures (CBMs; Deno, 1985, 2003). A CBM is a simple, efficient, and technically adequate measure used by educators to track student growth (Stecker et al., 2005; Petersen & Stoddard, 2018).

**Curriculum-Based Measures**

Deno (2003) explained that CBMs were created to bridge the gap between the measurement of student achievement and student instruction by allowing teachers to gather valid data on their students and base their instruction on these data. He further outlined the following essential characteristics of a CBM: standard measurement tasks, prescriptive stimulus materials, clear procedures for administration and scoring, performance sampling, multiple equivalent samples, time efficiency, easily taught procedures, and technical adequacy, which includes evidence of reliability and validity – validity being of utmost importance.

**Validity of Curriculum-Based Measures**

Messick (1989) defined validity as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment” (p. 13). Validity, in the case of CBMs and their measurement of reading, refers to their ability to adequately assess and recommend further action based on sufficient evidence and rationale. Such evidence and rationale can be gathered using specific evidence of four validity types: predictive validity, concurrent validity, content validity, and construct validity (Cronbach & Meehl, 1955; Petersen & Stoddard, 2018). For a CBM measuring reading ability, predictive validity can provide evidence that the measure adequately identifies young students who are truly at risk (sensitivity
and specificity) and determines how a student’s reading will either grow, remain unchanged, or decrease over time. Concurrent validity can provide evidence that the measure is comparable in validity to a measure that has already been established as having sufficient evidence. Content validity can provide evidence that the measure’s test items accurately represent what elements of decoding and language it is designed to measure. Construct validity can provide evidence that an assessment is measuring the construct of interest. A reading assessment that has construct validity will have evidence on how well the measure defines and operationalizes decoding and language so that it reflects the theory and academic expectations.

**Construct Validity**

Construct validity is of paramount importance. If a test does not have strong evidence of construct validity, then other sources of validity such as predictive and concurrent are less meaningful. There are several CBMs that focus on word level reading with strong evidence of construct validity. However, many current approaches to reading comprehension CBMs appear to have strong concurrent and predictive evidence of validity, yet their construct validity is questionable (Petersen & Stoddard, 2018). This limited evidence of construct validity negatively affects all aspects of the technical adequacy of a CBM, particularly prescriptive stimulus materials and performance sampling requirements. For example, if the prescriptive stimulus materials used by the CBM are not representative of the level of reading expected of the student or are not equivalent to the materials typical of the student’s grade level, then the CBM lacks adequate construct validity. Additionally, if the CBM does not accurately sample the performance of students to inform instruction, the CBM again lacks adequate construct validity.

**Current Curriculum Based Measures and Construct Validity of Reading Comprehension.** Current CBMs in the general education system used to identify early reading
comprehension difficulties in young students have two major flaws in their construct validity; (a) current widely-used assessments do not focus on language comprehension at all for young students (e.g., Aimsweb, Dynamic Indicators of Basic Early Literacy Skills [DIBELS]), which is a major component of the simple view of reading that must be measured, and (b) most CBMs used to measure reading comprehension do not sufficiently meet all of the psychometric requirements of a CBM as outlined by Deno (2003). These widely used CBMs rely on three approaches to measure reading comprehension: mazing, reading fluency, and number of words retold (Aimsweb, 2018; Good & Kaminski, 2002).

**Mazing.** Mazing refers to an assessment task wherein a student reads a passage that contains missing words. When they encounter the missing word, they are given multiple word options and are asked to pick which word they think is most appropriate for the sentence. A study by Muijselaar et al. (2017) investigated the CBM-Maze test to determine what maze assessment tasks measure and if these tasks offer useful information about a student’s reading comprehension compared to a reading comprehension assessment (i.e., Gates-MacGinite Reading Comprehension Test). It was found that while the CBM-Maze test correlated to the Gates-MacGinite test, indicating a degree of concurrent validity, it “relies more on code-related skills than on language comprehension skills” (p. 128). Additionally, when reading comprehension skills are measured by maze tasks, lower-level reading comprehension skills are primarily measured. If an element of a CBM such as mazing cannot give an educator reliable information about their students’ grade level reading comprehension skills, it would not meet the primary purpose of a CBM and should therefore be reconsidered.

**Reading Fluency.** Reading fluency, or oral reading fluency (ORF), is commonly operationalized as an assessment task wherein a student is asked to read a passage aloud and the
rate and accuracy of the student’s oral reading is measured. Reading fluency has been used to indirectly measure reading comprehension, yet when defined as rate and accuracy via commonly administered curriculum-based measures of oral reading fluency (CBM-Rs), it is primarily a measure of decoding/word-level reading. Yet, even as a decoding/word-level reading progress monitoring tool, it has weak psychometric evidence. In a systematic review by Ardoin et al. (2013) of recommendations and research regarding CBM-Rs, 102 instructional manuals, chapters, and articles were examined to determine whether the use of CBM-Rs is based in sufficient evidence. The authors found that “there are no studies to suggest that an individual student’s progress can be accurately determined” (p. 14) and the results of CBM-Rs should be interpreted cautiously.

**Counting Words From Retell.** In this assessment procedure, a student is asked to read a passage, retell what they read, and the number of words in the student’s retell is counted. Notably absent is the examination of the student’s understanding of the content they read. As reading comprehension necessarily involves language comprehension, a measure of the number of words retold would not determine a student’s understanding of a reading passage and therefore would not reveal a complete picture of a student’s reading comprehension skills (Cao & Kim, 2021). In a review of the psychometric properties of various retell instruments, Reed (2011) found that the DIBELS retell task, which counts the number of words in retell, did not give educators enough information about their students’ reading comprehension skills and therefore gave them little confidence in their students’ scores on the test.

In a study of 1,518 first-grade students examining efficacy of reading fluency and retell tasks, Riedel (2007) found that the reading fluency task was a better predictor of a student’s reading comprehension than the retell task. Given the previous discussion of the questionable
validity for measuring reading comprehension for a reading fluency task, results indicating that
the retell task is a poorer measure of reading comprehension should draw concern.

The CUBED Narrative Language Measures Listening and Reading Subtests

The CUBED Narrative Language Measures (NLM) is an oral language and reading
comprehension assessment that measures a student’s ability to understand and use written and
oral academic language. This test relies on the direct measurement of oral and written language
rather than peripheral measures such as fluency, mazing, and word counting to measure students’
reading comprehension skills. During the assessment, the students are instructed to listen to or
read aloud a brief story and then retell that story. The NLM consists of two subtests, the NLM
Listening and the NLM Reading, which use parallel narratives to test oral language, writing,
reading comprehension, and decoding fluency (reading fluency). The change of name of reading
fluency to decoding fluency is intentional to demonstrate that decoding is the only skill measured
in reading fluency tasks and should therefore not refer to a student’s overall reading skills.

The NLM Listening subtest is currently available for the assessment of five grade levels
(preschool, kindergarten, first, second, and third grade) and the NLM Reading subtest is
currently available for the assessment of three grade levels (first, second, and third grade). To the
extent possible, these levels are based on developmental patterns and curriculum expectations for
narrative structure, language complexity reflective of narration and exposition, and decoding
skills.

Both subtests assess a student’s oral language or reading comprehension and reading
fluency using personal-themed narratives. Personal-themed narratives were chosen for their
relevance to young children’s lives and because they require students to produce more complex
forms of language discourse. However, the complexity of the NLM narratives varies based on
grade level as students’ stories are expected to be less complex and organized at lower grade levels.

These narratives have been designed based on the conventions of curriculum-based measures (Deno, 2003). The narratives in the NLM have been created to be sampled repeatedly over time to inform educator decisions regarding language needs. Both the NLM Listening and the NLM Reading have one to two forms for fall, winter, and spring that can be administered at each grade (NLM Listening: preschool to third grade, NLM Reading: first to third grade) to sample student performance (performance sampling). To do so, the NLM has alternate forms with consistent lengths, Lexile ratings, language complexity, and narrative structure (multiple equivalent forms). The NLM narratives are key in ensuring that its assessment tasks are standardized (standardized assessment tasks). Additionally, digital prompts and booklets are included for examiners to follow to make procedures of administration clear and further standardize the administration (clear procedures for administration).

Each NLM story is constructed following precise narrative structure guidelines. Narrative structure, or story grammar, is comprised of specific structural elements. The story grammar in the NLM stories follows a pattern outlined by Stein and Glenn (1979), which tends to match mainstream, academic expectations. Story grammar elements are combined to form episodes in a narrative. The key story grammar elements of a basic episode include a problem, an attempt to solve the problem, and a consequence. NLM Preschool stories have very basic story grammar, with a main character, setting information with an activity and a location, a problem, the character’s feeling about the problem, an attempt to solve the problem, a consequence resulting from the attempt to solve the problem, and an ending. The NLM kindergarten through third grade stories have the inclusion of a plan that is stated right before the character’s attempt to
solve the problem, and end feelings describing how the character feels at the end of the story. Additional complexity is embedded in second and third grade stories in the form of multiple episodes, including a complication in which the first attempt to solve the problem does not work.

A narrative, like any other type of discourse, is also built from language complexity. Language complexity used in narration can be evaluated according to the extent that more complex and precise structures and vocabulary are present. A story can be produced using very basic vocabulary and grammar, but more successful, academically related narratives involve complex sentence structures with a clear time sequence and causal connections between events. Language complexity also increases with age. Average sentence length increases, and syntax becomes more complex. Language features such as openings, dialogue, adjectives, and temporal and adversative conjunctions (then, but) are in an earlier emerging stage. Adverbs, temporal subordinate clauses (e.g., after he got scared, he ran away), and multiple subordination are later developing language features. The second and third grade NLM stories have several linguistic features that mark high level oral language capabilities such as additional subordinate clauses, adverbs, adjectives, and low frequency vocabulary words.

The above guidelines were helpful in crafting parallel stories that are socially appropriate and cross-culturally relevant. Criteria for story grammar complexity and language complexity were initially drafted after a thorough review of the research on narrative developmental expectations and preliminary normative data from approximately 1000 preschool, kindergarten, first, second, and third grade students. Several empirical studies brought to light the extent to which typically developing students and students with language impairment could improve narrative structure and language complexity after narrative-based language instruction and intervention. Concurrently, a thorough review of the Common Core State Standards and other
state standards further shaped passage criteria, increasing language complexity and influencing the stories to reflect both narrative and expository language properties. Lexile indices primarily derived from the Common Core State Standards were calculated for each story, and editing took place to structure each passage until a Lexile range varying by only 10 to 20 points was achieved. In the final stages of creating stories for the NLM, each story underwent careful editing by linguists, speech-language pathologists, and the their authors. For the NLM Listening measure, feedback was elicited in field trials with teachers and speech-language pathologists in several states. Awkward wording and any additional editing suggestions were noted, and the passages were revised, processed through the strict a-priori development criteria, and molded to yield the appropriate Lexile index. For the NLM Reading, all 25 within grade-level stories were administered to several first, second, third, fourth, and fifth grade students. The number of correct words read in one minute was documented, as was accuracy, specific word errors, prosody, and the total time required to read the entire passage. Variability within and across students was noted, and consistently outlying stories were identified and revised. Revised stories were then administered to naïve students and again assessed for outlying status.

Each NLM Listening and NLM Reading story, within grade level, has nearly the exact same number of words (only varying from two to five words). Each story within grade level also consistently has the omission or inclusion of an exact, pre-specified number of story grammar elements, character name references, episodes, single pre-noun adjectives, double pre-noun adjectives, pre-verb adverbs, post-verb adverbs, adverbs of likelihood, non-overlapping tier-2 vocabulary words, the temporal conjunction then, the adversative conjunction but, the continuative conjunction so then, left-branching temporal subordinate clauses with the subordinating conjunction when, left-branching temporal subordinate clauses with the
subordinating conjunction after, right-branching causal subordinate clauses with the subordinating conjunction because, right-branching causal subordinate clauses with the subordinating conjunction so that, left-branching adverbial subordinate clauses with the subordinating conjunction although, adjectival (relative) subordinate clauses using that or who, two or more subordinate clauses in the same utterance not associated with dialogue, dialogue, and appositives with and without relative clauses. Each story is written with ‘universal’ themes in mind that can transcend culture (e.g., getting hurt), with specific accounts that are plausible to have occurred in the lives of preschool and school-age children in the U.S. and other countries.

Each NLM Listening and NLM Reading story has also been equated using the Lexile index. The first several utterances of the first grade, second grade, and third grade NLM Reading stories have a lower Lexile index to accommodate beginning readers, but the entire story still matches the Lexile index of the NLM Listening passages.

**Statement of the Problem**

As 63%-69% of fourth and eighth grade students currently read below a proficient level (NAEP, 2022) and current CBMs do not measure necessary constructs of reading comprehension, a CBM that does measure these necessary constructs is needed. Because the NLM directly measures oral and written language comprehension, it clearly measures the construct of interest. This patent construct validity is a considerable strength of the NLM measures. However, research is needed to provide evidence of concurrent validity and predictive validity including sensitivity and specificity.
Statement of the Purpose

The purpose of this study was to examine the validity of the CUBED NLM Listening and NLM Reading subtests, specifically to examine their concurrent, and predictive validity, including sensitivity and specificity.

Research Questions

This study addressed the following research questions or research hypotheses:

1. To what extent do the CUBED NLM universal screening and progress monitoring subtests adequately correlate with valid academic oral and written language assessments (Curriculum-Based Assessment for Writing, a narrative language sample, expository language, state standardized reading comprehension tests, and a general oral language outcome) for students in preschool through third grade?

2. To what extent do the CUBED NLM universal screening and progress monitoring subtests predict summative reading assessment results (MAPS, PAWS – state summative reading assessment) and yield adequate sensitivity and specificity for students in kindergarten through third grade?

Method

Participants

For evidence of concurrent validity for the CUBED NLM, data from 1,146 preschool through third grade students were examined. Sixty nine percent of the participants were white, 13% were Hispanic, 5% were African American, 3% were Native American, 1% were Asian, and 3% were other. Six percent of the participants had a language disorder.
For evidence of predictive validity for the CUBED NLM, we examined data from 1,512 kindergarten through third grade students. For sensitivity and specificity analyses we included 71 third grade students attending two different elementary schools in Wyoming.

**Procedures**

**Concurrent Validity**

Evidence of validity can be derived by examining the relationship between the CUBED and the results of other assessments administered at approximately the same time. This examination is designed to demonstrate the extent that the CUBED matches comparable, previously validated instruments in student performances (Bachman, 1990; McNamara, 2000). This convergent evidence of validity, often represented as *concurrent criterion-related validity* (concurrent validity for short), is highly dependent on the existence of valid, external tests.

**Predictive Validity**

Evidence of validity can be derived by examining how well the CUBED relates to future performance on other tests designed to measure similar constructs (Anastasi & Urbina, 1997). Because results from the CUBED in the fall were used to place students into language and decoding interventions, the predictive validity analyses are confounded by having used the predictive criteria to help prevent failure on the criterion measures. Clearer conclusions may have been possible if this biasing relationship were not present, although given earlier research, failure to intervene would have had negative ethical implications and a negative impact on student performance.

**Sensitivity and Specificity.** Sensitivity and specificity were evaluated by examining how well the CUBED identified students who were performing at or below grade level according to the end of year state of Wyoming Proficiency Assessments for Wyoming Students (PAWS).
Measures

For concurrent validity, we compared the CUBED NLM Listening retell highest score to scores from several criterion measures of language (scores for a curriculum-based assessment for writing, narrative language sample [Frog, Where Are You?], expository language, Clinical Evaluation of Language Fundamentals Preschool [CELF-P], and Renfrew Bus Story). We also compared the CUBED NLM Listening retell scores to the Measures of Academic Progress (MAP) assessment. The CUBED NLM Listening retell scores were calculated by converting z-scores to scaled scores, and then summing those scaled scores to reflect decoding, language, and reading constructs. To demonstrate the predictive validity of the CUBED, we studied its relationship to the MAP and Wyoming PAWS reading assessments.

Curriculum-Based Assessment for Writing

To obtain authentic samples of students’ writing, we obtained school district administered writing artifacts. Students were administered a District Writing Assessment (DWA) that measures the adequacy of a student’s writing ability using a grading rubric specific to their grade level. The grading rubric examines seven aspects of writing: purpose/voice, organization, ideas, sentence fluency, word choice, conventions, and presentation. Each aspect of writing is scored on a scale of 0–4, with a total of 28 points possible. The pretest writing prompt administered to all participants in the treatment and control groups was expository-related and the posttest prompt, also administered to all students in the treatment and control groups, was narrative focused. For this study, the DWAs from winter (pretest) and spring (posttest) were administered and scored by the teachers, who received training from their school district on the administration and scoring of the test. However, both teachers were unaware that the DWA would be used as an outcome measure in this study (Petersen et al., 2020).
Narrative Language Sample (Frog, Where Are You?)

Narrative language retell samples have long been considered an authentic and valid measure of a student’s ability to produce and understand complex academic language (Westby, 1985). The wordless storybook *Frog, Where Are You?* (Mayer, 1969) is commonly used to elicit a narrative retell sample and using that book to elicit a language sample has strong evidence of validity. In this study, examiners read a *Frog, Where Are You?* script from the Systemic Analysis of Language Transcripts (SALT) manual (Miller & Iglesias, 2012) while showing the child the corresponding pictures from the wordless picture book. Following the modeled story, the child was given the wordless picture book and asked to retell the story. Each retell sample was audio-recorded, transcribed, segmented, and analyzed by trained research assistants using the SALT software (Miller & Iglesias, 2012).

Expository Language

Expository language was assessed by speech-language pathologists (SLP) or by paraprofessionals under SLP supervision using a set of experimental materials and procedures, which were designed to be a type of criterion-referenced assessment of informational text comprehension and production. Following a similar retell format as the NLM Listening subtest, the examiners asked each student to listen to information about an unusual animal and then retell that information. Students were also asked three factual questions about the passage. Scoring was completed in real time while the child was retelling the information and answering questions. Main ideas and key details were scored according to a 2-point scale based on their clarity and completeness. The three factual questions were also scored on a 2-point scale. There was a combined total of 26 points possible. Two forms of the expository measure were designed for use in this study; one about sea pigs was used at pretest in the winter, and one about bird-
eating spiders was used at posttest in the spring. The information in each expository measure was designed to be particularly obscure information that students would likely have not been exposed to previously. These expository retells took approximately one minute to administer and score.

**Clinical Evaluation of Language Fundamentals Preschool 2 (CELF-P 2)**

The Clinical Evaluation of Language Fundamentals-Preschool 2 (CELF-P 2) in English (Wiig et al., 2004) is an omnibus norm-referenced test designed to measure general language abilities. For this study, four subtests were administered: Word Structure, Sentence Structure, Expressive Vocabulary, and Following Directions. The CELF-P 2 has been found to be valid and reliable with test-retest reliability for both subtests and composite scores across all ages ranging from 0.78 to 0.98. Additionally, the internal consistency reliability of the CELF-P 2 ranged from 0.79 to 0.97, demonstrating acceptable scores across subtests and excellent composite scores.

**Renfrew Bus Story**

The Renfrew Bus Story is a screening assessment used in conjunction with other tools to guide diagnosis (Hayward et al., 2008) of language disorders. The Bus Story measures the student’s ability to retell relevant story information. This assessment is administered individually, the student listens as the test administrator reads the story aloud then they are asked to retell the story and given appropriate cueing. The student’s retell abilities are interpreted based on the information retold, the length of their sentences, the complexity of their sentences, and the level of cueing needed.

**Measures of Academic Progress (MAP)**

The Measures of Academic Progress (MAP) is a norm-referenced, grade independent, computer-adaptive instrument designed to be a general outcome measure and progress monitoring tool in the areas of reading and mathematics (NWEA, 2019). An overall reading
Rasch Unit (RIT) score as analyzed in this study. For the MAP reading assessment, students were asked to silently read information and then answer multiple-choice questions. The test consistently adapts to each student’s response by delivering a more difficult question if the student answered correctly or an easier question if the student answered incorrectly. The reading portion of the MAP consists of 40–50 multiple-choice questions evaluating three different areas of reading: literature, informational text, and foundational skills and vocabulary. Additionally, MAP Language and Writing, MAP Literature and Informational, and MAP Vocabulary Acquisition and Use and MAP Use and Functions were used as criterion measures. The MAP Language and Writing scores measured a student’s writing conventions (i.e., spelling and punctuation) and their grammar, language, and usage skills. The MAP Literature and Informational measures the student’s ability to identify structure and important ideas within literature or a given text. The MAP Vocabulary Acquisition and Use measures a student’s ability to know the meaning of multiple vocabulary words. The Vocabulary Use and Functions measures a student’s ability to use references and context clues to understand and apply vocabulary.

**Wyoming Proficiency Assessments for Wyoming Students (PAWS)**

Wyoming PAWS assesses the reading, writing, math, and science progress of students from third to eighth grade and eleventh grade (Educational Testing Service, 2015). The PAWS test is aligned with the Wyoming Content and Performance Standards. For the purposes of this study, only the reading progress results were examined. Sensitivity in this case represents the extent to which the CUBED accurately identified students who were at risk on PAWS, and specificity refers to the extent to which the CUBED accurately identified students who were not
at risk on PAWS. The PAWS assessment has evidence of validity (Educational Testing Service, 2015).

**Results**

The scores of preschool through third grade students on the CUBED NLM Listening and Reading subtests were evaluated to determine the subtests’ concurrent and predictive validity. First, concurrent validity was examined by calculating the correlations between the NLM subtests and valid academic oral and written language assessments. Pearson’s product moment correlation coefficients are presented and discussed below. Second, predictive validity of the NLM subtests on summative reading assessments was evaluated using regression and ROC analyses including sensitivity and specificity. In addition, evidence of adequate sensitivity and specificity for students in kindergarten through third grade was found using $R^2$ values and is also discussed and shown below.

**Concurrent Validity**

In the following analyses, we compared the CUBED NLM Listening and Reading retell highest score to scores from several criterion measures of oral and written language using Pearson’s product moment correlation coefficients (Tables 1 & 2). The correlations provided evidence of concurrent validity. Positive correlation coefficients were interpreted as follows: .20 to .29 were considered weak, coefficients ranging from .30 to .39 as moderate, coefficients ranging from .40 to .69 as strong, and coefficients at or above .70 were considered very strong. Student performance reported in the studies used to evidence criterion-related validity most likely did not represent the full range of performance on the criterion measures. This restricted range in test scores tends to underestimate the correlations (Hunter & Schmidt, 1990; Henriksson
& Wolming, 1998; Thorndike, 1949). Thus, coefficients have been corrected for range and attenuation. Corrected and uncorrected coefficients were reported.

**Predictive Validity**

We used regression and receiver operating characteristic (ROC) analyses to examine the predictive validity, including the sensitivity and specificity. In Tables 3, 4, and 5 we reported $R^2$ coefficients of determination to indicate the extent to which the NLM raw scores collected in the fall with 1,512 kindergarten through third grade students were predictive of the MAP assessment in the winter. $R^2$ interpretation is highly dependent upon the testing context. Therefore, for predicting reading, $R^2$ values above .10 (10% accounted variance) were considered meaningful. For diagnostic instruments, such as norm-referenced standardized language assessments, sensitivity and specificity at or above 80% were considered generally considered acceptable as universal screeners such as the CUBED are often expected to be lower in sensitivity and specificity. Sensitivity and specificity provide evidence of both predictive validity and construct validity. Sensitivity and specificity were analyzed through logistic regression and receiver operator characteristic (ROC) analyses in Table 6.

**Discussion**

The purpose of this study was to examine the validity of the CUBED NLM Listening and NLM Reading subtests, specifically their concurrent and predictive validity, including their sensitivity and specificity. In this study we included 1,146 preschool through third-grade participants to examine concurrent validity, 1,515 kindergarten through third-grade participants to examine predictive validity, and 71 third grade students to examine sensitivity and specificity. We administered the NLM Listening in the fall to all participants and the NLM Reading in the Fall to first through third grade participants. We then administered several criterion measures of
oral language, writing, and reading comprehension in the fall and winter and the PAWS state of Wyoming reading assessment in the spring.

**Concurrent Validity**

Results indicate that the NLM Listening and NLM Reading both correlated moderately to strongly with the criterion measures used for concurrent evidence of validity (Research Question 1). When the NLM scores are correlated with the MAP writing score, the relationship is strong. This indicates that even though the NLM Listening measures the integration of the comprehension and production of complex, oral academic language, it reflects a student’s writing performance. These data show evidence of the relationship between oral language and written language. Thus, even though writing samples are not routinely collected or analyzed using the NLM, NLM data can provide evidence of a student’s writing ability.

When the NLM Listening and Reading subtests were compared to the MAP fall language writing assessment, the NLM had evidence of strong correlation. This result indicates that the NLM Listening can identify students with potential writing difficulties as early as kindergarten. This finding is an important step in identifying academic needs in young students and providing them with the necessary supports to create improved educational outcomes.

When comparing the NLM Listening to *Frog, Where Are You?* (Meyer, 1969), the NLM Listening has evidence of moderate to strong correlations with narrative language retells including the episode complexity, story grammar, total number of words, number of different words, subordination index, and words per minute for all grades. There are strong correlations with kindergarten through third grade students for mean length of utterance and total number of utterances, but those correlations were only moderate with the preschool students. *Frog, Where Are You?* is a wordless picture book that has been used to elicit oral narrative language samples
from thousands of children across multiple studies (Miller & Iglesias, 2012). There is considerable evidence that *Frog, Where Are You?* can validly elicit language samples (Heilman et al., 2010). Finding that the NLM Listening correlates moderately to strongly with outcome measures from the *Frog, Where Are You?* strengthens the evidence of validity of the NLM. Although the correlations were moderate to strong between the NLM Listening and the *Frog, Where Are You?* outcome measures, both approaches elicit narrative retell language samples and the relationship between those language samples might have been higher if (a) the NLM Listening had used pictures to elicit the narrative retell language sample and (b) the NLM Listening had included multiple episodes in the model narrative, and (c) had used less complex, academic language. Because of these differences, the correlations were likely attenuated.

The NLM Listening and Reading subtests also correlate strongly with a general reading outcome, as measured using the MAP RIT score. This correlation is particularly important, because it provides evidence that oral language as measured using the NLM can provide information on a student’s reading comprehension performance. Thus, the NLM can provide information on a student’s potential reading comprehension performance even with students in preschool and kindergarten who cannot yet decode. By identifying students early for comprehension difficulty, we can provide early intervention that may result in better academic outcomes.

The NLM Listening and Reading subtests are also found to correlate strongly with expository language. This indicates that the NLM may operate as a general outcome measure for complex language ability across multiple contexts. This finding is important because reading comprehension is measured using narrative and expository language which could contain any
narrative theme or content knowledge. A curriculum-based measure should be able to provide information on oral language ability that transcends those contextual factors (Kamhi, 2014).

Results also indicated that the NLM Listening and Reading subtests correlated weakly with specific measures of vocabulary words (MAP Vocabulary Acquisition and Use) yet correlated strongly with the MAP vocabulary assessment that measured a student’s ability to infer the meaning of vocabulary words from context (MAP Vocabulary Use and Functions). It is not surprising that the NLM did not correlate strongly with the vocabulary assessment that measured specific words, given that it is highly unlikely that the Tier-2 words used in the MAP assessment were the same Tier-2 words used in the NLM. However, the NLM was primarily designed to measure a student’s general ability to learn vocabulary from context, thus the strong relationship between the MAP vocabulary assessment that measures inferential word learning and the NLM provide considerable evidence of validity for the NLM.

Finally, the concurrent evidence of validity for the NLM Listening when compared to the CELF-P and the Renfrew Bus Story was very strong for the preschool students examined in this study. These findings indicate that the NLM Listening can provide general language information, even though the NLM Listening takes less than 2 minutes to administer. Current curriculum-based general outcome measures for oral or written language are efficient (they usually take less than 5 minutes to administer), yet they lack considerable evidence of validity (Petersen & Stoddard, 2018). It has been particularly difficult to develop a measure of language that meets all Deno’s parameters for curriculum-based measures, especially navigating the tension between efficiency and validity. The results from this study indicate that the NLM Listening, which is efficient, strongly correlates with general language measures.
Predictive Validity

Results indicated that the NLM Listening and the NLM Reading separately and combined, accounted for meaningful variance in the criterion measures (Research Question 2). The fall NLM Listening results were a stronger predictor of reading comprehension on the winter MAP assessment than the NLM Reading. This difference could be due to the higher average Lexile rating for the NLM Reading passages (610-800L) compared to the MAP reading passages (450-725L; Northwest Evaluation Association, 2013). The Lexile rating is a measure of the complexity of a reading passage, and a higher rating on the NLM Reading would indicate that it may have been more difficult for the students to decode and therefore understand when compared to the MAP reading passages. This higher Lexile rating for the NLM Reading is equivalent to the Lexile rating for the NLM Listening, with the only difference between the Reading and Listening forms being that one is read by the student whereas the other is read to the student by the examiner. This higher Lexile rating for the NLM Reading and NLM Listening was purposeful. The purpose of the NLM is to primarily measure how well a student can understand complex academic language. Thus, for the younger grades, the NLM Listening may be a better indicator of comprehension because word recognition is less of a factor that can impede comprehension. It is also interesting to note that both the NLM Listening and NLM Reading were weak predictors of the comprehension of expository text. This is not surprising given that the NLM assesses narrative discourse. Future research should examine the extent to which embedding expository information in the NLM narratives increases the predictive validity of the NLM for expository text comprehension.

The results of this study also indicated that the combined NLM Listening and NLM Reading had 77% sensitivity and 100% specificity for the PAWS end-of-year assessment for the
third-grade students and had 76% sensitivity and 70% specificity for the MAP Winter Reading composite (RIT) score, which is fair to excellent. Although not excellent, this predictive validity is notable given that only language comprehension-related variables from the CUBED were used to predict future reading performance. Additional subtests of the CUBED which measure word-recognition, including reading fluency, would likely add to the predictive validity of the CUBED given that a combination of word recognition and language comprehension would account for both components of the simple view of reading.

**Validity of the Narrative Language Measures**

It has historically been difficult to validly measure oral language within the parameters outlined for a CBM by Deno (2003), such as standard measurement tasks, prescriptive stimulus materials, clear procedures for administration and scoring, performance sampling, multiple equivalent samples, time efficiency, easily taught procedures, and reliability and validity, and as a result, most CBMs do not focus on oral language with young students (see Aimsweb, 2018; Good & Kaminski, 2002). Yet oral language is a key variable in the simple view of reading, where word recognition multiplied by language comprehension equals reading comprehension (Gough & Tunmer, 1986). Without tools to assess oral language in young students, educators must wait until a student can decode written language before they obtain information on comprehension, and students are not expected to decode complex academic language until around the third grade. Therefore, it is not until third or fourth grade that educators begin to have an awareness of which students have difficulty understanding complex, academic language. Waiting until third or fourth grade is unnecessary because oral language can be measured beginning in preschool. The NLM allows for the measurement of oral language at an early age,
thereby facilitating the early identification and prevention of future reading comprehension difficulty (Lyon, 2003).

Even when the measurement of comprehension does take place, proxy measures that do not directly measure language comprehension and therefore lack construct validity are typically used (e.g., mazing, reading fluency, and number of words retold; Aimsweb, 2018; Good & Kaminski, 2002). Using the NLM, examiners elicit and score narrative language samples, which reflect the construct of academic language. This study provided additional evidence that the NLM can directly measure and predict oral language and reading outcomes, indicating that it could possibly be used as a universal screener and progress monitoring curriculum-based measure, especially since it meets all the other parameters outlined by Deno (2003), including efficiency.

The criterion “gold standard” measures used in this study have considerable evidence of validity for reflecting child language ability (i.e., writing samples; narrative language samples elicited using *Frog, Where Are You?*; expository language samples; omnibus norm-referenced language testing, and state- and district- administered summative reading assessments). The NLM correlated moderately to strongly with each of those “gold standard” measures, indicating that the NLM also measures the same or a similar construct. Deno (2003) noted that in addition to being valid and reliable, a curriculum-based measure should be time efficient. The NLM takes approximately 2-3 minutes to administer and score whereas the criterion measures or “gold standard” measures of language included in this study take considerably longer to administer and score. It is crucial for a CBM for oral language to be efficient, as 60-80% of students across the U.S. are not meeting curriculum standards for reading (NAEP, 2022), and many of those students are struggling with the comprehension of complex, academic language, therefore their progress
in oral language and reading comprehension must be monitored frequently to ensure that the proper interventions are being provided and can be adjusted as necessary.

Clinical Implications

This study indicates that the NLM may be a useful tool to reduce the percentage of students reading below grade-level. Educators can use this tool to identify students with reading difficulty earlier in their education and subsequently provide early intervention as well as identify students who will struggle on formative and summative reading comprehension assessments in the future.

Additionally, the ability of educators to directly examine a student’s oral language with the NLM as opposed to usage of proxy measures allows educators to obtain direct information regarding a student’s language and treatment targets.

Limitations and Future Research

This study only includes sensitivity and specificity data for 71 third-grade students from Wyoming and only predicted performance within grade level. Most participants were white and Hispanic, with little data for Black, Native American, Asian, and other races or ethnicities.

It is important to note that the NLM Listening and the NLM Reading subtests are only two subtests from a battery of subtests in the CUBED assessment. Future research should examine the extent to which the NLM coupled with other subtests including embedded expository discourse, factual and inferential reasoning measures, reading fluency, phonemic awareness, letter sounds, irregular words, and measures providing information on a student’s ability to decode multiple syllable types will accurately assess and predict future reading comprehension.
References


https://doi.org/10.1097/01.tld.0000318937.39301.76


https://doi.org/10.1044/2014_LSHSS-13-0063


Evaluation Association.


Tables

Table 1

*Unattenuated/Corrected (Uncorrected) Correlation Coefficients Between the Fall NLM Listening and Language-Related Criterion Measures, Evidencing Concurrent Validity*

<table>
<thead>
<tr>
<th>Criterion Measure</th>
<th>n</th>
<th>r (Preschool)</th>
<th>r (Grades K-3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Writing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLM Listening Total Retell and MAP Fall Language Writing</td>
<td>1085</td>
<td>.56* (.48)</td>
<td></td>
</tr>
<tr>
<td>NLM Total Retell and <em>Frog, Where Are You?</em> Language Sample</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLM Total Retell And Episode Complexity</td>
<td>50</td>
<td>.69* (.52)</td>
<td></td>
</tr>
<tr>
<td>NLM Total Retell And Story Grammar</td>
<td>50</td>
<td>.67* (.52)</td>
<td></td>
</tr>
<tr>
<td>NLM Total Retell and FWAY NDW&lt;sup&gt;a&lt;/sup&gt;</td>
<td>112</td>
<td>.55 (.44)</td>
<td>.61* (.54)</td>
</tr>
<tr>
<td>NLM Total Retell and FWAY NDW&lt;sup&gt;b&lt;/sup&gt;</td>
<td>166</td>
<td>.65* (.58)</td>
<td></td>
</tr>
<tr>
<td>NLM Total Retell and FWAY TNW&lt;sup&gt;a&lt;/sup&gt;</td>
<td>112</td>
<td>.51 (.41)</td>
<td>.58* (.49)</td>
</tr>
<tr>
<td>NLM Total Retell and FWAY TNW&lt;sup&gt;b&lt;/sup&gt;</td>
<td>166</td>
<td>.25 (.21)</td>
<td></td>
</tr>
<tr>
<td>NLM Total Retell and FWAY MLU&lt;sup&gt;a&lt;/sup&gt;</td>
<td>112</td>
<td>.35 (.28)</td>
<td>.60* (.53)</td>
</tr>
<tr>
<td>Criterion Measure</td>
<td>n</td>
<td>(r) (Preschool)</td>
<td>(r) (Grades K-3)</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>----</td>
<td>-------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>NLM Total Retell and (FWAY) MLU(^b)</td>
<td>166</td>
<td>.49(^*) (.44)</td>
<td></td>
</tr>
<tr>
<td>NLM Total Retell and (FWAY) TNU(^a)</td>
<td>112</td>
<td>.36 (.29)</td>
<td>.58(^*) (.52)</td>
</tr>
<tr>
<td>NLM Total Retell and (FWAY) TNU(^b)</td>
<td>166</td>
<td>.47(^*) (.42)</td>
<td></td>
</tr>
<tr>
<td>NLM Total Retell and (FWAY) SI(^b)</td>
<td>166</td>
<td>.24 (.21)</td>
<td></td>
</tr>
<tr>
<td>NLM Total Retell and (FWAY) WPM(^b)</td>
<td>166</td>
<td>.48(^*) (.41)</td>
<td></td>
</tr>
</tbody>
</table>

**NLM Productivity and Complexity Measures and \(Frog, Where Are You?\) Language Sample**

<table>
<thead>
<tr>
<th>Criterion Measure</th>
<th>n</th>
<th>(r) (Preschool)</th>
<th>(r) (Grades K-3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLM NDW and (FWAY) NDW(^b)</td>
<td>166</td>
<td>.65(^*) (.58)</td>
<td></td>
</tr>
<tr>
<td>NLM TNW and (FWAY) TNW(^b)</td>
<td>166</td>
<td>.27(^*) (.23)</td>
<td></td>
</tr>
<tr>
<td>NLM MLU and (FWAY) MLU(^b)</td>
<td>166</td>
<td>.63(^*) (.56)</td>
<td></td>
</tr>
<tr>
<td>NLM TNU and (FWAY) TNU(^b)</td>
<td>166</td>
<td>.54(^*) (.48)</td>
<td></td>
</tr>
<tr>
<td>NLM SI and (FWAY) SI(^b)</td>
<td>166</td>
<td>.44(^*) (.39)</td>
<td></td>
</tr>
<tr>
<td>NLM WPM and (FWAY) WPM(^b)</td>
<td>166</td>
<td>.63(^*) (.54)</td>
<td></td>
</tr>
<tr>
<td>Criterion Measure</td>
<td>n</td>
<td>( r ) (Preschool)</td>
<td>( r ) (Grades K-3)</td>
</tr>
<tr>
<td>-------------------------------------------------------</td>
<td>----</td>
<td>----------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>General Reading Outcome</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP RIT(^a)</td>
<td>1086</td>
<td>.57* (.49)</td>
<td></td>
</tr>
<tr>
<td>Expository Language</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information Retell(^a)</td>
<td>917</td>
<td>.42 (.32)</td>
<td>.68* (.50)</td>
</tr>
<tr>
<td>NLM Total Retell and Researcher Expository Retell</td>
<td>85</td>
<td>.51* (.41)</td>
<td></td>
</tr>
<tr>
<td>NLM Total Retell and MAP Expository (Literature</td>
<td>548</td>
<td></td>
<td>.29 (.25)</td>
</tr>
<tr>
<td>Informational)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary Use</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLM Total Retell and MAP Vocabulary Acquisition and</td>
<td>548</td>
<td>.26 (.22)</td>
<td></td>
</tr>
<tr>
<td>Use</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLM Total Retell and MAP Vocabulary Use and Functions</td>
<td>1085</td>
<td>.53* (.46)</td>
<td></td>
</tr>
<tr>
<td>Clinical Evaluation of Language Fundamentals Preschool (CELF-P)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLM Total Retell and CELF-P Core(^a)</td>
<td>62</td>
<td>.59* (.47)</td>
<td></td>
</tr>
<tr>
<td>Criterion Measure</td>
<td>n</td>
<td>$r$ (Preschool)</td>
<td>$r$ (Grades K-3)</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>---</td>
<td>----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>NLM Total Retell and Information</td>
<td>5</td>
<td>.95* (.88)</td>
<td></td>
</tr>
</tbody>
</table>

**Renfrew Bus Story**

*Note.* a study 1, b study 2, *strong correlation, NLM = Narrative Language Measures; FWAY = *Frog, Where Are You?*; NDW = Number Different Words; TNW = Total Number of Words; MLU = Mean Length of Utterance; TNU = Total Number of Utterances; SI = Subordination Index; WPM = Words per Minute. All correlations were significant, $p < .05$. For corrected coefficients, estimates for *Frog, Where Are You?* reliability were estimated to be .79 unless specified in Heilmann et al. (2008; NDW = .79, NTW = .70, WPM = .74). Reading Street reading comprehension end of unit test reliability data and reliability for the district writing assessment were not available. We estimated .70 for those tests. Expository was .70. Bus Story reliability for Information was .79 according to the Bus Story Manual. Average reliability for the CELF was .80. Average reliability for the nonword repletion task was .89. Average reliability for the NLM Listening was .95. Average reliability for the MAP assessment was .80.
Table 2

*Unattenuated/Corrected (Uncorrected) Correlation Coefficients Between the Fall NLM Reading and Language-Related Criterion Measures, Evidencing Concurrent Validity*

<table>
<thead>
<tr>
<th>Criterion Measure</th>
<th>n</th>
<th>( r ) (Preschool)</th>
<th>( r ) (Grades K-3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Writing</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLM Listening Total Retell and MAP Fall Language Writing</td>
<td>1085</td>
<td>.56* (.48)</td>
<td></td>
</tr>
<tr>
<td><strong>General Reading Outcome</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP RIT</td>
<td>1086</td>
<td>.57* (.49)</td>
<td></td>
</tr>
<tr>
<td><strong>Expository Language</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLM Total Retell and Researcher Expository Retell</td>
<td>85</td>
<td>.51* (.41)</td>
<td></td>
</tr>
<tr>
<td>NLM Total Retell and MAP Expository (Literature Informational)</td>
<td>548</td>
<td>.29 (.25)</td>
<td></td>
</tr>
<tr>
<td><strong>Vocabulary Use</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLM Total Retell and MAP Vocabulary Acquisition and Use</td>
<td>548</td>
<td>.26 (.22)</td>
<td></td>
</tr>
<tr>
<td>NLM Total Retell and MAP Vocabulary Use and Functions</td>
<td>1085</td>
<td>.53* (.46)</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* *strong correlation, MAP reliability was estimated as .80. Researcher Expository reliability was estimated at .70. NLM = Narrative Language Measures, MAP = Measures of Academic Progress.
Table 3

Regression Analyses Using Fall NLM Listening Raw Scores to Predict Winter MAP Performance Evidencing Predictive Validity

<table>
<thead>
<tr>
<th>Criterion Measure</th>
<th>n</th>
<th>$R^2$ Grades K-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLM Fall Predicting MAP RIT Winter (Study 1)</td>
<td>1064</td>
<td>.25</td>
</tr>
<tr>
<td>MAP Language and Writing</td>
<td>1063</td>
<td>.28 (.24)</td>
</tr>
<tr>
<td>MAP Information and Literature</td>
<td>540</td>
<td>.09 (.08)</td>
</tr>
<tr>
<td>MAP Vocabulary Use and Functions</td>
<td>1063</td>
<td>.28 (.24)</td>
</tr>
<tr>
<td>MAP Vocabulary Acquisition and Use</td>
<td>522</td>
<td>.20 (.17)</td>
</tr>
</tbody>
</table>

Note. All $R^2$ values were significant, $p < .001$; NLM = Narrative Language Measures, Measures of Academic Progress fall administration; RIT = Rasch Unit.
Table 4

Regression Analyses Using Fall NLM Reading Raw Scores to Predict Winter MAP Performance Evidencing Predictive Validity

<table>
<thead>
<tr>
<th>Criterion Measure</th>
<th>n</th>
<th>R^2 (Grades 1-3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLM Fall Predicting MAP RIT Winter (study 1)</td>
<td>737</td>
<td>.15</td>
</tr>
<tr>
<td>MAP Language and Writing</td>
<td>731</td>
<td>.13(.11)</td>
</tr>
<tr>
<td>MAP Information and Literature</td>
<td>218</td>
<td>.09(.08)</td>
</tr>
<tr>
<td>MAP Vocabulary Use and Functions</td>
<td>731</td>
<td>.13(.11)</td>
</tr>
<tr>
<td>MAP Vocabulary Acquisition and Use</td>
<td>520</td>
<td>.07(.06)</td>
</tr>
</tbody>
</table>

Note. All R^2 values were significant, p < .001; NLM = Narrative Language Measures, MAP = Measures of Academic Progress, RIT = Rasch Unit.
### Table 5
Correlated (Uncorrelated) Correlations Between the Fall NLM Listening and Reading Scaled Score Composite and Winter MAP

<table>
<thead>
<tr>
<th>Criterion Measure</th>
<th>n</th>
<th>$R^2$ (Grades K-3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measures of Academic Progress (MAP) Winter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winter RIT Score</td>
<td>681</td>
<td>.21(.18)</td>
</tr>
<tr>
<td>Winter MAP Language and Writing</td>
<td>685</td>
<td>.21(.18)</td>
</tr>
<tr>
<td>Winter MAP Information and Literature</td>
<td>218</td>
<td>.09(.08)</td>
</tr>
<tr>
<td>Winter MAP Vocabulary Use and Functions</td>
<td>731</td>
<td>.13(.11)</td>
</tr>
<tr>
<td>Winter MAP Vocabulary Acquisition Use</td>
<td>520</td>
<td>.07(.06)</td>
</tr>
</tbody>
</table>

*Note. All $R^2$ values were significant, $p < .001$; RIT = Rasch Unit Progress, NLM = Narrative Language Measures, MAP = Measures of Academic Progress.*
Table 6

*Sensitivity and Specificity of NLM Listening and NLM Reading Combined Third Grade Fall Assessment for End of Year PAWS Reading Assessment and Middle of Year MAP Reading Assessment*

<table>
<thead>
<tr>
<th>End of Year Reading Criterion Measure</th>
<th>N</th>
<th>AUC</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLM Reading and NLM Listening Combined Predicting PAWS Spring</td>
<td>71</td>
<td>.86</td>
<td>77%</td>
<td>100%</td>
</tr>
<tr>
<td>NLM Reading and NLM Listening Combined Predicting MAP Winter</td>
<td>68</td>
<td>.76</td>
<td>76%</td>
<td>70%</td>
</tr>
</tbody>
</table>

*Note.* PAWS Spring = State of Wyoming Reading Assessment administered in March; Poor performance on PAWS = at or below basic. MAP Winter = Measures of Academic administered in January; Poor performance on MAP = at or below basic.
APPENDIX A

Annotated Bibliography


**Objective:** This article is a systematic review of CBM-R literature that is often used to support the use of the CBM-R for progress monitoring and decision making.

**Method:** This article identified 171 journal articles, instructional manuals, and chapters using research databases and online search engines. The authors chose 102 sources for this review using study criteria.

**Results:** The review conducted in this article indicated that the use of the CBM-R for program evaluation and progress monitoring lacks evidence.

**Relevance to current work:** The CBM-R is an example of a curriculum-based measure of reading comprehension that primarily uses reading fluency, a measure that only accounts for decoding or word-recognition abilities. This thesis presents the argument that a CBM used to measure reading comprehension should measure both components of the simple view of reading and if a measure lacks either component, it does not have adequate construct validity.


[https://doi.org/10.1016/j.edurev.2020.100375](https://doi.org/10.1016/j.edurev.2020.100375)
**Objective:** This meta-analysis sought to determine the use of retell in measuring reading comprehension in comparison to other measures of reading comprehension, such as word retold.

**Method:** The data of 23 studies were found using electronic databases (ERIC, PsycINFO, and Sociological Abstracts) and the study’s criteria. These studies were then analyzed.

**Results:** This study found that measuring structural elements or quality of retell is more accurate than counting number of words retold when measuring a student’s reading comprehension ability.

**Relevance to current work:** This thesis proposes that curriculum-based measures using number of words retold lack adequate construct validity as this kind of measure is not an accurate indicator of a student’s reading comprehension abilities.


**Objective:** The object of this study was to examine and compare the language abilities of children with poor reading comprehension, children with poor decoding ability, and typical readers to test the simple view of reading hypothesis.

**Method:** Performance data on measures of language comprehension and phonological processing at the eighth-grade level from 57 students with poor reading comprehension, 27 students with poor decoding ability, and 98 typical readers was examined.
Results: The results of this study indicate that students that have difficulty with reading comprehension can be classified by their ability to comprehend language, decode, or both, thus supporting the simple view of reading hypothesis.

Relevance to current work: The simple view of reading hypothesis forms the basis of how the validity of reading comprehension measures is determined. The findings from this study determines that the simple view of reading has psychometric strength and can reasonably be used as the basis of this discussion of validity.


This article details ways in which the simple view of reading has impacted the clinical and research fields of speech-language pathology. It explains the ways in which the simple view of reading has been expanded on to add important information regarding potential nuances encountered when distinguishing between language comprehension.

Relevance to current work: The simple view of reading is a foundational concept that explains why current curriculum-based measures are lacking validity in their measure of reading comprehension and other measures such as the CUBED Narrative Language Measures may fill this gap in assessments.


This article is focused on the concept of validity in relation to psychological testing. The authors describe the concept of validity as it pertains to these tests and defines validity in four ways: concurrent validity, predative validity, content validity, and construct validity.
Although the context of this article is psychology, the authors define these terms in general research method terms.

**Relevance to current work:** The topic of validity is foundational to the current work. The definitions provided by the authors of this article aid in defining the terms predictive validity and concurrent validity accurately so that the research base for this work’s questions can be adequately constructed.

https://doi.org/10.1207/s1532799xssr0601_03

**Objective:** The objective of this study was to evaluate the ways in which linguistic comprehension and phonological abilities impact the development of reading comprehension.

**Method:** The study included longitudinal data of 141 students from first grade to third grade. Measures for phonological awareness, listening comprehension, decoding, and reading comprehension were evaluated at the end of first grade and third grade.

**Results:** Results of the study indicated that word decoding, vocabulary, and listening comprehension all influence the development of reading comprehension. This suggests that each is an important component when considering a student’s reading comprehension abilities.

**Relevance to current work:** The current work is based on the simple view of reading, that both language comprehension and decoding are important elements of reading comprehension. The findings of this study help support this view.

This article describes the importance of regular assessment of students’ academic performance and the need for data from these assessments to be valid and reliable. Validity and reliability are vital so that educators can accurately implement necessary adjustments in their teaching to meet the students’ needs. Deno explains the curriculum-based measures (CBMs) can provide this valid and reliable data.

**Relevance to current work:** This article helps make the case for the development and use of CBMs such as the CUBED Narrative Language Measures. Additionally, it provides a framework to evaluate CBMs such as the CUBED NLM and determine whether it has greater validity than commonly used CBMs.


This article adds to the discussion of curriculum-based measures (CBMs) found in Deno (1986). The author explains the growth in research regarding CBMs, leading to a discussion of best practices in creating and using CBMs. This includes a list of components that a CBM should include to be a valid and reliable measure (technical adequacy, standard measure tasks, prescriptive stimulus materials, administration and scoring guidelines, performance sampling, multiple equivalent samples, time efficiency, and ease of teaching).

**Relevance to current work:** The components outlined in this article add an additional, necessary framework for evaluating the validity of CBMs such as the CUBED Narrative Language Measures.

This article discusses the role of screening processes in identifying students who are at risk for various learning disabilities. The overall goal of the article is to explain what is understood about screenings and their importance and then to propose their effective implementation in, for example, response-to-intervention models in schools.

**Relevance to current work:** The discussion in this article regarding screening processes and their role in identifying at risk students gives support to the current work’s assertion that valid screeners are necessary to identify students with reading difficulties early in their education and provide them with necessary interventions.


This article explains the role of decoding in reading comprehension, support for and against decoding in its importance when considering reading comprehension ability. The authors propose the simple view of reading as a perspective on the issue of the role of decoding. It proposes that while decoding is likely an important element of reading comprehension, it alone is not sufficient. It must be considered in addition to listening or language comprehension to provide a more complete data on reading comprehension.

**Relevance to current work:** The simple view of reading as proposed in this article is the foundation this work. The current work proposes that, based on the simple view of reading, decoding and language comprehension must be evaluated in reading
comprehension measures and current curriculum-based measures commonly used are not valid as only decoding ability is measured.


This article reviews evidence that supports the simple view of reading. It outlines ways in which various assessments used to assess reading comprehension use decoding but lack measures of language comprehension. It then details evidence that indicates the importance of measuring language comprehension, especially in elementary education, to accurately identify students with reading difficulties.

**Relevance to current work:** The role of language comprehension in reading comprehension and the lack of it in commonly used measures is key to the current work. The review of evidence found in this article supports the assertion that more assessments such as the CUBED Narrative Language Measures should be created and evaluated to gather data on reading comprehension starting in early education.


**Objective:** The purpose of this study was to evaluate the influence of decoding and language comprehension on reading comprehension. The goal was to examine the claims made by the simple view of reading.
**Method:** In this study, 757 students from third grade through fifth grade participated in measures of decoding and reading comprehension to measure each and its relation to reading comprehension.

**Results:** The study found that both components of the simple view of reading, language comprehension and decoding, accounted for nearly all variance in reading comprehension. That is, each student found to have reading difficulties also had difficulties with either or both components.

**Relevance to current work:** The current work is supported by the assertion that both language comprehension and decoding are important when evaluating reading comprehension and any measure of reading that does not adequately measure both components is not a valid measure. This study supports this assertion by indicating that decoding and language comprehension account for almost all variance in reading comprehension.


This purpose of this article is to outline the development of reading comprehension and common difficulties encounter by students in early education while developing reading comprehension. The authors emphasize that difficulties in reading comprehension can be overcome but such difficulties must be identified early in education in increase the rate of success. Educators must have valid tools to identify reading difficulties and data from assessments to inform their teaching.

**Relevance to current work:** The current work highlights that based on current data, most students in fourth and eighth grade do not meet grade level benchmarks for
reading. This article supports the notion that progress monitoring tools should be
developed to measure reading comprehension in early elementary grades.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of

https://doi.org/10.3102/0013189x018002005

This purpose of this article is to explain the concept of validity in the context of scientific
research. The authors provide various definitions of validity to identify the overall
purpose of validity in research and how this purpose can be sued to create a framework
for research.

**Relevance to current work:** A discussion of validity and its importance in
research is used heavily in the current work. This article provides concise definitions of
validity that aid in this discussion and create clarity for the case of evaluating the validity
of measures such as the CUBED Narrative Language Measures.


https://doi.org/10.1080/10888438.2016.1263994

**Objective:** The objective of this study was to examine code-related (fluency, decoding)
and language comprehension (vocabulary, listening comprehension) tasks in the CBM-
Maze test to determine whether the test adequately measures both decoding and language
comprehension.

**Method:** This study used sample of 274 students in grades 4, 7, and 9 (92 fourth
graders, 90 seventh graders, and 92 ninth graders). The test scores of this sample using
the CBM-Maze test and the Gates-MacGinite Reading Comprehension Test were
examined using a multigroup structural regression mode to evaluate the different code-related and language comprehension demands.

**Results:** The results of this study determined that the CBM-Maze test relies on decoding more than language comprehension skills with consistency across all grade groups.

**Relevance to current work:** The current work asserts that that mazing relies more on decoding than language comprehension and therefore cannot be classified as a reading comprehension measure based on the simple view of reading hypothesis.


The objective of this article is to examine the various psychometric needed in assessments of oral and written language to ensure validity and reliability. The authors explained various approaches to creating progress monitoring tools that are educator friendly. They then outlined various requirements to ensure the psychometric integrity of these tools.

**Relevance to current work:** The requirements outlined in this paper help build the case within the current work as to why commonly used progress monitoring tools do not meet psychometric ideals for measuring oral and written language.

**Objective:** This paper is a narrative synthesis reviewing the psychometric properties of 11 different retell instruments that are used to measure reading comprehension in K-12 students, including the DIBELS retell task.

**Method:** The retell instruments reviewed were found using ancestral search of references used in a separate synthesis by the author and using the databases of various test publishers.

**Results:** When reviewing the DIBELS retell task, which uses number of words in a retell as a measurement, it was found that the task did not give educators enough information about the students’ reading comprehension abilities.

**Relevance to current work:** This thesis states that a CBM of reading comprehension, such as the DIBELS retell task, should both be valid in its construct and give educators accurate information about student performance to appropriately tailor instruction to their needs (Deno, 2003). Because the DIBELS retell task does neither, it lacks validity as a measure of reading comprehension.


**Objective:** This study examined the relationship between the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), reading comprehension, and vocabulary skills of students.

**Method:** The performances of 1,518 first-grade students on DIBELS (including the Oral Reading Fluency subtest), the Group Reading Assessment and Diagnostic Evaluation (GRA+DE), and TerraNova Reading subtest.
**Results:** Among other findings, the study found that reading fluency was a better predictor of reading comprehension ability than retell tasks that counted number of words retold.

**Relevance to current work:** This thesis presents data that reading fluency is not a valid measure of reading comprehension. The finding in this study that words retold is weaker in its prediction of reading comprehension than reading fluency further makes the case that measure that utilize words retold lack sufficient construct validity.


This article reviews the efficacy of using curriculum-based measures (CBMs) as tools for improving the achievement of students. The authors describe various studies wherein CBMs were used to monitor student progress in reading and mathematics and in what ways data from these measures were implemented into educators’ teaching. The authors concluded that general improvement was found in student achievement when CBMs were used.

**Relevance to current work:** The current work argues that progress monitoring tools such as CBMs should be used to identify at risk students so that they can receive necessary services as early in their education as possible. The conclusion of this article aids the current work in this argument by supporting the claim that CBMs lead to increased academic achievement and are therefore important measures to develop and evaluate.
APPENDIX B

Consent/Institutional Review Board Approval Letter

UNIVERSITY OF WYOMING

Vice President for Research & Economic Development
1000 E. University Avenue. Department 3355 • Room 305/308. Old Main • Laramie, WY 82071
(307) 766-5353 • (307) 766-5320 • fax (307) 766-2608 • www.uwyo.edu/research

March 10, 2023

Douglas Peterson
Professor
Communication Disorders
University of Wyoming

Alisa Therkildsen
Student
Communication Disorders
University of Wyoming

Carlee Flanagan
Project Manager
Communication Disorders
University of Wyoming

Protocol #20230310DP03512

Re: IRB Proposal "Examining the Validity and Reliability of the CUBED assessment as a Universal Screening and Progress Monitoring Tool"

Dear Douglas, Alisa, and Carlee:

The proposal referenced above qualifies for exempt review and is approved as one that would not involve more than minimal risk to participants. Our exempt review and approval will be reported to the IRB at their next convened meeting on March 23, 2023.

Any significant change(s) in the research/project protocol(s) from what was approved should be submitted to the IRB (Protocol Update Form) for review and approval prior to initiating any change. Further information and the forms referenced above may be accessed at the "Human Subjects" link on the Office of Research and Economic Development website: http://www.uwyo.edu/research/human-subjects/index.html. Please note that exempt protocols are approved for a maximum of three years. If your study extends beyond three years, or beyond the duration that is approved in your protocol form, please be sure to submit an update before expiration to extend the duration. If you are not able to submit the update in time, you will need to submit a new exemption request for the project.

You may proceed with the project/research and we wish you luck in the endeavor. Please feel free to call me if you have any questions.

Sincerely,

Nichole Person

Nichole Person
Coord., Research Compliance, Research Office
On behalf of the Chairman.
Institutional Review Board