5-2-2012

Language Profile and Performances on Math Assessments for Children with Mild Intellectual Disabilities

Katherine T. Rhodes
Georgia State University

LANGUAGE PROFILE AND PERFORMANCES ON MATH ASSESSMENTS FOR CHILDREN WITH MILD INTELLECTUAL DISABILITIES

by

KATHERINE T. RHODES

Under the Direction of Rose A. Sevcik

ABSTRACT

It has been assumed that mathematics testing indicates the development of mathematics concepts, but the linguistic demands of assessment have not been evaluated, especially for children with mild intellectual disabilities. 244 children (grades 2 – 5) were recruited from a larger reading intervention study. Using a multilevel longitudinal SEM model, baseline and post-intervention time points were examined for the contribution of *item linguistic complexity*, *child language skills*, and their potential interaction in predicting item level mathematics assessment performance. *Item linguistic complexity* was an important, stable, and negative predictor of mathematics achievement with children's language skills significantly and positively predicting mathematics achievement. The interaction between *item linguistic complexity* and *language skills* was significant though not stable across time. Following intervention, children with higher language skills performed better on linguistically complex mathematics items. Mathematics achievement may be related to an interaction between children's language skills and the linguistic demands of the tests themselves.

INDEX WORDS: Mild Intellectual Disability (MID), Linguistic complexity, Child language skills, Mathematics assessment performance, KeyMath-Revised

LANGUAGE PROFILE AND PERFORMANCES ON MATH ASSESSMENTS FOR CHILDREN WITH MILD INTELLECTUAL DISABILITIES

by

KATHERINE T. RHODES

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Arts

in the College of Arts and Sciences

Georgia State University

2012

LANGUAGE PROFILE AND PERFORMANCES ON MATH ASSESSMENTS FOR CHILDREN WITH MILD INTELLECTUAL DISABILITIES

by

KATHERINE T. RHODES

Committee Chair: Rose A. Sevcik

Committee: Robin Morris

MaryAnn Romski

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2012

ACKNOWLEDGMENTS

This research would not have been possible without the participation of students, teachers, and schools included in the Reading Intervention Study. The study research staff and senior researchers also dedicated countless hours to accomplish this research. Thank you to my Thesis Committee, Rose Sevcik, Robin Morris, and MaryAnn Romski, for their guidance and support. I also would like to extend a special thank you to Jack Barille, Chris Henrich, Bradley Goodnight, Karie Gaska, and Lee Sanford.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: LITERATURE REVIEW

General mathematics skills are an important aspect of successful daily living. School-age children in the United States are regularly tested for mathematics proficiency, and the results of these tests are used to inform curriculum development and intervention efforts for those students who are not performing at grade level. Although there is concern about poor mathematics achievement in the overall U.S. population of school age children, those children with mild to moderate disabilities have the largest achievement gap compared to their peers without disabilities. Children with mild intellectual disabilities represent a large portion of the U.S. population of children with developmental disabilities, and their specific mathematics achievement profile is an area in need of additional research to design targeted interventions. While it often has been assumed that mathematics testing results indicate poor development of mathematics concepts, the linguistic demands of auditory processing and verbal working memory have not been substantially evaluated as potentially confounding assessment effects for children with mild intellectual disability. Intervention efforts targeting only mathematics concepts, without attention to the language skills needed to interpret assessment demands, may be ineffective for children with certain cognitive-linguistic profiles.

To address this area of concern, the current study sought to examine the role of cognitive linguistic skills and mathematics assessment performance in children with mild intellectual disabilities. The purpose of this research was to characterize the relationship between the linguistic demands of mathematics assessments and language skills of school-age children with mild intellectual disabilities in predicting the likelihood that these children would be able to correctly answer mathematics assessment items.

Mathematics Achievement in the United States

Basic mathematics skills are essential in all aspects of independent living (e.g., navigating personal finances, measuring distances, planning events and manipulating schedules, etc.). Mathematics achievement from Kindergarten to postgraduate levels is a focus in the objectives of the U.S. Department of Education because it is also vital to achievement in the sciences (STEM Education Coalition, 2000). Through agencies like the National Science Foundation and the Institute of Education Sciences, the U.S. government has attempted to monitor and improve national education trends in mathematics. However, many children in the U.S. still fail to achieve grade level proficiency.

The most recent National Assessment of Educational Progress (NAEP) National Report Card indicated that a significant percentage of students were below grade level proficiency in the 2011 national sample. The NAEP reported that across all students tested, 60% of fourth graders in the United States were below grade level proficiency in mathematics (NCES, 2011). As grade level increases, the trend of mathematics achievement is worse, with 65% of 8th graders performing below grade level proficiency (NCES, 2011). Although 12th grade was not assessed during the most recent (2011) National Report Card, in the 2009 assessment 74% of 12th graders were performing below grade level proficiency in mathematics (IES, 2010). Of the school-age children tested in 2011, approximately 11% of the 4th grade students tested and 10% of the 8th grade students tested were children identified as having one or more disabling conditions (e.g., hearing difficulty, visual difficulty, learning disability, mild intellectual disability; NCES, 2011).

Disparities in U.S. Mathematics Achievement for Children with Disabilities

Although students with disabilities represented a relatively small proportion of the overall sample of children tested, they evidence the largest achievement gap in mathematics when

compared to children without disabilities. Among those students who were labeled with a disability, 83% of 4th graders and a devastating 91% of 8th graders were below grade level proficiency (NCES, 2011; note that the 2009 results indicated that 93% of 12th graders with disabilities were performing below grade level proficiency, IES, 2010). In contrast, those children who were not identified as having disabilities did not evidence the same extreme achievement problems; 57% of 4th grade students and 62% of 8th grade students who did not have a disability performed below grade level proficiency (NCES, 2011).

Although the national achievement statistics often dichotomize disability, educational research has provided some clarity as to the achievement trends of children with mild disabilities (including learning disabilities, emotional-behavioral disorders, and mild intellectual disabilities). In analyzing the mathematics achievement literature for children with mild disabilities, Parmar, Frazita, and Cawley (1996) identified three troubling achievement trends, (1) the contemporary age-to-grade performance of children with mild disabilities is similar to what it was in the 1930s, (2) the rate of growth for children with mild disabilities is approximately one year of grade level achievement for every two years of school, and (3) by the end of secondary school, children with mild disabilities reach only the 5th or 6th grade level of mathematics knowledge with competencies in arithmetic computation and no competencies in problem solving and applications.

Historically, students with disabilities have been excluded from large scale achievement studies, because their participation requires special testing accommodations. Since 1996, the NAEP has been working to ensure more testing accommodations for students with disabilities (U.S. Department of Education, 2009; including access to a dictionary, breaks during testing, receiving cues to stay on task, receiving directions read aloud, receiving extra time for testing,

access to large-print or magnification devices, testing in small groups, etc.). However, despite increased efforts towards inclusion, exclusion of persons with disabilities was still an issue in the 2011 National Report Card: students could be excluded by their schools (not by independent assessments with the NAEP examiners) on the basis of labels of severe cognitive impairments, concerns that testing accommodations would necessitate multiple days of testing, or concerns that students would require non-permitted testing accommodations (U.S. Department of Education, 2009).

Those students who were included in the most recent NAEP testing most likely represented children with mild to moderate disabilities; however, specific characterization of these students is difficult because "disability" is not specifically described in terms of overall functioning or diagnosis. The NAEP dichotomizes disability in terms of those students who were labeled with one or more disabilities, as evidenced by IEPs and other school records, and those students who were not labeled as having a disability (NCES, 2011). A dichotomized treatment of disability does not allow for the characterization of specific types of disabilities (e.g., learning disabilities, intellectual disabilities, autism spectrum disorders, hearing impairments, visual impairments, etc.); it allows only for generalizations about the achievement trends of some portion of those children receiving special education who were selected for test participation by their schools. Understanding national mathematics learning and achievement profiles relative to specific disability diagnoses is an area in need of additional research.

Among developmental disabilities, intellectual disability is the most common and ranks first among conditions causing major limitations in activity in the U.S. (CDC, 1996), but children with intellectual disabilities have not been included in much of the developmental research on mathematics achievement difficulty to date. While prevalence estimates vary by region, target

age, definitions and measurements of disability, and overall study methodology, there seems to be some agreement that U.S. national prevalence for mild intellectual disability in school-age children is between 1% and 3% (Roeleveld, Zielhuis, & Gabreels, 1997). The 1991, 1996, and 2000 MADDSP reports have consistently estimated that mild intellectual disability accounts for approximately two thirds of children with intellectual disability (Bhasin, Brocksen, Avchen, & Van Naarden Braun, 2006; Boyle, Yeargin-Allsopp, Doernberg, Holmgreen, Murphy, & Schendel, 1996). Thus, a large percentage of the children with disabilities reported in national achievement testing are most likely children with mild intellectual disability.

Challenges In Assessing Mathematics Achievement

Many popular mathematics assessments (e.g., the Kaufman Assessment Battery for Children, KABC, Kaufman & Kaufman, 1985; the Wechsler Individual Achievement Test, WIAT, Psychological Corporation, 1992; the KeyMath-Revised Diagnostic Inventory of Essential Mathematics, KM-R, Connolly, 1988) have been criticized for their lack of content validity for use with children who have mild disabilities (Parmar, Fazita, & Cawley, 1996). These assessments often fail to provide balanced coverage of mathematic concepts, focusing largely on arithmetic computation and not on strategy and problem solving. The content reflected in assessments is also not always relevant to the curriculum emphasized at the classroom level or in students' IEPs, and thus, testing recommendations may have little practical relevance to educational placement, curriculum design, and instructional strategies.

The mathematics achievement tests most commonly used in the U.S. rely on dichotomous (right/wrong) scoring systems for evaluating students' responses, but a dichotomized scoring system does not allow for the characterization of cognitive features contributing to mathematics difficulty. A right/wrong scoring system allows for the identification of students who are

struggling to provide correct answers in various mathematics content areas, but it does not provide insight as to why. A failure to provide a correct answer on a mathematics assessment item can be the result of any number of errors, (1) an error in understanding what one is being asked to do, (2) an error in selecting the correct approach to a mathematics question or the correct operation to a mathematics problem, (3) a computational error in correctly completing a mathematics operation, or (4) an error in reporting the correct answer one has derived (Goodstein, Kahn, & Cawley, 1976). While some types of errors are more indicative of difficulty with mathematics knowledge or skill (which could be the result of cognitive difficulties with mathematics concepts or instructional shortcomings in relaying mathematics information), other errors could be indicative of more general cognitive, linguistic, or even motor difficulties.

Few studies have addressed the specific pattern of mathematics assessment errors for children with mild intellectual disabilities (MID). Error analysis studies conducted with other populations of children with mild disabilities (including learning disabilities and emotional behavioral disabilities but excluding MID) indicate that most of the mathematics errors are being made in (1) correctly interpreting the instructions and linguistic demands of the question, and (2) selecting the correct operation and approach to the question (Parmar, 1992). In Parmar's (1992) study with 31 children with learning disabilities or emotional-behavioral disorders aged 8 to 14 years, reading questions aloud was not sufficient to help students with disabilities identify and remember key features of the problems, break problems down into steps, or integrate the relevant steps of a solution. These students also struggled with selecting operations that were appropriate for solving particular problems and matching operations to arithmetic symbols, even though they were able to correctly carry out computations once the appropriate operations had been identified. Other obstacles for these students included (1) difficulty self-monitoring and self-

correction, (2) difficulty selectively attending to relevant information and suppressing extraneous information, and (3) difficulty with concentration for prolonged periods without prompts (Parmar, 1992). The error patterns evidenced for other students with mild disabilities would seem to suggest that even with the provision of testing accommodations, students with MID may not be able to access the linguistic demands of mathematics assessment items.

The fact that testing accommodations are needed for the inclusion of children with (mild to moderate) disabilities in national achievement testing highlights the major issues of validity in using standardized mathematics assessments with populations of children who have disabilities. Children with intellectual disabilities are routinely assessed with measures that were designed and normed using typically developing children. The vast majority of testing accommodations provided to children during the 2011 NAEP study involved reading test questions and/or directions aloud. These allowed testing accommodations are in place, not to reduce the mathematics content demands of the assessment questions, but to enable students with disabilities to access testing instructions and the meanings of the questions themselves. Such accommodations are derived from students' individualized educational plans (IEPs), and they are used across standardized testing scenarios (NCES, 2011). However, even with the provision of these testing accommodations, the mathematics achievement disparity of children with disabilities remains extreme. This pattern may speak to the linguistic demands of mathematics assessments.

Language Difficulties and Assessment Demands for Children with MID

Language is commonly understood as some combination of skills in the areas of syntax, morphology, vocabulary (including expressive and receptive vocabulary knowledge), semantics, and pragmatics (Bloom & Lahey, 1978). Broader features of cognitive functioning such as

auditory processing (specifically phonological awareness), social knowledge, working memory, and executive functioning also may be incorporated to understand and measure language. For the purposes of this study, language was of interest insofar as language was used in direct mathematics testing situations. Child language abilities were considered in terms of syntax, morphology, vocabulary, and semantics.

For children with mild intellectual disabilities, language functioning is often a significant impairment for overall functioning. Miller, Chapman, and MacKenzi (1981) reported that for approximately 50% of children with intellectual disability, language comprehension and/or production is significantly below the level of general cognitive functioning. The auditory processing tasks of attending to relevant cues, discriminating between similar and different cues, organizing and categorizing cues, storing and retrieving cues, and synthesizing linguistic information (both simultaneously and sequentially) may all represent significant challenges for children with intellectual disabilities (Owens, Metz, & Haas, 2007).

As mathematics questions become more complex along the dimensions of syntax, morphology, vocabulary, and semantics, they may become more difficult for children with MID to answer. While it often has been assumed that mathematics testing results indicate poor development of mathematics concepts for these children, the linguistic demands of auditory processing and verbal working memory have not been substantially explored as potentially confounding assessment effects.

Measures of mathematics ability often rely heavily on language as the primary modality of question delivery and response delivery. Paper/pencil or verbal formats are common for mathematics assessments, while less linguistically demanding formats are more rare (e.g., using manipulatives, pictoral displays, or pointing/gesturing formats; Parmar, Frazita, & Cawley,

1996). Though accommodations may be provided to aid children with intellectual disabilities in completing mathematics assessments (reading questions aloud, repeating verbal stimuli, and allowing extra time and prompting), the auditory language processing demands of assessment may still present significant challenges for children with MID. Language-heavy assessments may unintentionally become measures of language ability, as opposed to measures of mathematics ability, when used with children who have language difficulties. The current study sought to characterize the relationship between children's language profiles, mathematics assessment items' linguistic complexity, and children's mathematics performance.

Research Questions and Hypotheses

Research Question 1.

What is the relationship between item linguistic complexity and children’s language skills in predicting the mathematics achievement of children with mild intellectual disability (MID)? It is expected that both item linguistic complexity and children's language skills are significant predictors of mathematics achievement.

Research Question 2.

Is there an association between mathematics performance and item linguistic complexity, and if so, is that association dependent upon a child’s language profile? It is expected that a significant interaction between item linguistic complexity and children's language skills predicts mathematics performance.

Research Question 3.

Is the relationship between item linguistic complexity and children's language skills stable over time in its prediction of mathematics achievement? It is expected that the interaction

between item linguistic complexity and children's language skills is not stable over time, such that intervention experiences can positively impact children's abilities to cope with linguistically complex mathematics items.

CHAPTER 2: METHODS

Participants

The participants were recruited for a reading intervention study designed to test the efficacy of reading programs for students with mild intellectual disability. The parent study spanned five years from 2005 to 2010 (Sevcik, 2005). Participants were selected using initial school-based referrals and then screened for additional inclusionary and exclusionary criteria.

Schools in the greater metro-Atlanta area referred children who were between the ages of 7 (at the end of the first grade) and 10 (at the end of fourth grade), met the state criteria for mild intellectual disability, and were eligible for special education services. Consent packets were sent home for parents to review, and participation was allowed for those students who returned completed consent forms.

Students were eligible for inclusion in the reading intervention study if they demonstrated difficulty in developing reading skills. Students were excluded from the study if they spoke English as a second language, demonstrated hearing impairment, demonstrated uncorrected vision impairment, or had a history of serious emotional and/or psychiatric disturbance based on school records. Recruitment also attempted to balance the sample across the sexes.

A final sample of 244 children, who completed one year of intervention, was selected for the current study from the reading intervention study sample. At baseline this sample ranged in age from 80 months to 147 months, with a mean age of 110.80 months (SD = 16.18). The overall sample grade level mean was 3.33 (SD = 1.14). These children represented two metro-Atlanta area counties, A (n = 78) and B (n = 166), and 12 schools. The mean PPVT language age of this sample was 4.80 years (SD = 1.63), and the mean IQ of this sample was 63.09 (SD = 9.40, Minimum = 37.00, Maximum = 87.00). Valid IQ scores were provided by participating schools

for 206 of the total 244 students participating. Missing data patterns were considered during subsequent data analysis. Approximately 62.20% of the sample was male (n = 158). The sample was racially and ethnically diverse (56.15% African American, 20.90% Caucasian, 16.39% Hispanic, 2.05% Asian, 4.10% Multiracial, and .41% Not reported).

Parents of students eligible for participation were asked to complete a family demographic questionnaire including information about parent education and income and information about child developmental and medical history. The information from this questionnaire was then coded using the Hollingshead Two Factor Index of Social Position in order to obtain numerical values for socioeconomic labels (Hollingshead, 1975). The mean family Hollingshead score was 30.28 (SD = 12.96), the mean level of education for mothers (n=226 respondents) was 12.70 years (SD = 3.03), and the mean level of education for fathers (n=155 respondents) was 12.60 years (SD = 3.63). Table 1 presents a breakdown of continuous descriptive variables for the overall sample and for each intervention group. Table 2 presents frequency data for discrete descriptive variables for the overall sample and for each intervention group.

Measures

Measures overview.

An assessment battery was selected by the parent study researchers to describe students' initial cognitive and linguistic profiles and to assess the outcomes of the interventions in areas of academic achievement and language skills. This assessment battery was administered at four time points, (1) at a baseline time point when students had received 0 hours of intervention instruction, (2) after 60 hours of instruction, (3) after 120 hours of instruction, and finally (4) at a follow-up time point one year after intervention ended.

Table 1

Continuous Demographic Variables N, Mean, and Range Overall and By Intervention Group

	Overall Sample			PHAB Group			RAVEO Group			Math Group		
		Mean						Mean			Mean	
	n	(SD)	Min - Max	n	Mean	Min - Max	n	(SD)	Min - Max	n	(SD)	Min - Max
Age (months)	244	110.80	80 - 147	87	112.14	84 - 144	80	116.36	86 - 144	77	103.49	80 - 147
		(16.18)			(15.38)			(16.10)			(14.53)	
PPVT Lang. Age	244	4.80	1.09 - 11.04	87	4.98	1.09 - 8.07	80	5.08	1.09 - 11.04	77	4.29	1.11 - 7.02
		(1.63)			(1.59)			(1.85)			(1.29)	
IQ	206	63.09	37 - 87	73	62.97	48 - 86	66	62.14	44 - 84	67	64.15	37 - 87
		(9.40)			(8.87)			(9.21)			(10.16)	
Grade Level	244	3.33	2 - 5	87	3.55	2 - 5	80	3.69	2 - 5	77	2.70	2 - 4
		(1.14)			(1.17)			(1.16)			(0.76)	
Hollingshead Scores												
Family Overall Score	224	30.28	8 - 66	82	29.85	8 - 66	70	32.49	10.5 - 63.5	72	28.61	8 - 58
		(12.96)			(12.08)			(13.23)			(13.52)	
Mother Ed.	226	12.70	0 - 19	79	13.20	6 - 19	74	13.23	3 - 18	73	11.62	0 - 18
		(3.03)			(2.51)			(2.68)			(3.57)	
Mother Ed. Score	226	4.40	1 - 7	79	4.58	1 - 7	74	4.65	1 - 7	73	3.96	1 - 7
		(1.41)			(1.26)			(1.29)			(1.58)	
Mother Occupation Score	220	3.29	1 - 9	78	3.10	1 - 9	71	3.49	1 - 8	71	3.28	1 - 8
		(2.44)			(2.33)			(2.62)			(2.40)	
Father Ed.	155	12.60	0 - 22	56	12.63	3 - 22	53	13.42	3 - 22	46	11.63	0 - 18
		(3.63)			(3.61)			(3.42)			(3.74)	

Father Ed. Score	155	4.31	1 - 7	56	4.32	1 - 7	53	4.62	1 - 7	46	3.93	1 - 7
		(1.52)			(1.51)			(1.44)			(1.57)	
Father Occupation Score	147	4.18	1 - 9	54	3.96	1 - 9	50	4.66	1 - 9	43	3.91	1 - 9
		(2.24)			(2.09)			(2.34)			(2.27)	
Mother Hollingshead Score	216	29.91	8 - 66	76	29.46	11 - 66	70	31.74	8 - 61	70	28.57	8 - 58
		(14.86)			(13.94)			(15.24)			(15.46)	
Father Hollingshead Score	142	34.22	8 - 66	53	33.17	8 - 66	47	37.72	13 - 66	42	31.62	8 - 66
		(14.03)			(13.53)			(14.25)			(13.96)	

Note. PPVT: Peabody Picture Vocabulary Test III Form A (Dunn & Dunn, 1997). Hollingshead Two Factor Index of Social Position (Hollingshead, 1975). IQ measures vary across schools and students.

Table 2

Non-Continuous Demographic Variables Frequency Data Overall and By Intervention Group

	Overall Frequency	PHAB Frequency	RAVEO Frequency	Math Frequency
IQ	244	87	80	77
Below 50	11	5	3	3
50 – 70	157	55	53	49
Above 70	38	13	10	15
Missing	38	14	14	10
Grade Level	244	87	80	77
2nd	80	24	19	37
3rd	54	15	13	26
4th	60	24	22	14
5th	50	24	26	0
Sex	244	87	80	77
Male	158	55	48	55
Female	86	32	32	22
Ethnicity	244	87	80	77
African American	137	52	44	41
Caucasian	51	14	22	15
Hispanic	40	12	11	17
Asian	5	2	1	2
Mixed	10	7	2	1
Not Reported	1	0	0	1
Schools	244	87	80	77
1	44	18	9	17
2	31	10	21	0
3	12	8	1	3
4	10	7	3	0
5	6	3	0	3
6	28	8	13	7
7	3	0	3	0
8	18	6	8	4
9	30	6	9	15
10	13	5	8	0
11	27	6	5	16
12	22	10	0	12
Counties	244	87	80	77
A	78	29	22	27
B	166	58	58	50

Note. For the purposes of sample characterization, IQ and grade level were characterized as discrete variables in this table; however, both were treated as continuous variables in subsequent analyses.

Mathematics achievement measure.

The KeyMath-Revised Inventory (KM-R; Connolly, 1988) is one of the most widely used mathematics assessments for children with disabilities receiving special education services. The KM-R was designed to assess students' basic math competencies in concepts, operations, and applications across a variety of math domains. The KM-R is diagnostic in the sense that it provides measurement of performance across specific areas of mathematics curriculum (e.g., numeration, addition, subtraction, geometry, problem solving, etc.), but it does not yield specific patterns of error analysis (e.g., incorrect algorithm selection, computational error, etc.; Goodstein, Kahn, & Cawley, 1976).

The KM-R was designed for students in grades K through 9, normed on 1,794 typically developing students 5 to 15 years of age. It consists of 13 subscales, each representing a major concentration of mathematics skills. For the purposes of the parent study and the current project, only six subscales of the KM-R (Form A) were administered to the students. The Numeration, Geometry, Addition, Subtraction, Measurement, and Time and Money subscales were selected to reflect the curriculum experiences of students with mild intellectual disability receiving special education in grades 1 to 5.

Questions on the KM-R are administered orally with minimal visual support from an illustration array, and student responses are provided orally. For example, an examiner might administer an item on the numeration subscale by saying, "How many children do you see in this picture," while the student observes a visual array depicting several children of varying sizes and orientations on a playground. A few items are administered using written mathematical symbols on operations subscales such as Addition or Subtraction.

Split half reliability coefficients for the KM-R assessment are dependent on subtest and grade level. For the numeration subtest, students in grades 1 through 5 of the normative sample all demonstrated reliability coefficients at or above .75. For the geometry subtest, students in grades 1 through 5 demonstrated reliability coefficients at or above .72. For the addition subtest, students in grades 1 through 5 demonstrated reliability coefficients at or above .56. For the subtraction subtest, grades 1 through 5 demonstrated reliability coefficients at or above .68. For the measurement subtest, students in grades 1 through 5 demonstrated reliability coefficients at or above .72. For the time and money subtest, students in grades 1 through 5 demonstrated reliability coefficients at or above .67 (Connolly, 1988).

Content validity for the KM-R was examined using essential math content to reflect curricula and national trends, consultations with numerous experts in mathematics education, and subdivision of the assessment into domains to reflect equal weighting among concepts. However, the content validity of the KM-R, when used with populations of children who have mild disabilities, has been called into question for failure to provide balanced coverage of mathematics concepts, overemphasis on computation and under emphasis on problem solving, and mismatch with students' special education classroom experiences and IEP goals (Parmar, Fazita, & Cawley, 1996).

Construct validity of the KM-R was examined using developmental stage progression analyses, reliability analyses, and convergent validity with the Comprehensive Test of Basic Skills (with an overall correlation of .66) and the Iowa Test of Basic Skills (with an overall correlation of .76; Connolly, 1988). However, Connolly (1988) did not provide empirical evidence for the proposed KM-R factor structure as a means of establishing construct validity, and the factor structure and construct validity of the KM-R and subsequent versions of the

KeyMath have been criticized by a number of researchers (Walker & Arnault, 1991; Williams, Fall, Eaves, Darch, & Woods-Groves, 2007; see Appendix A).

The KM-R assessment was not timed. Each subscale was administered until students reached a ceiling with three consecutive incorrect responses. At the item level, correct responses to each item are recorded as '1' and incorrect responses are recorded as '0'. Raw scores can be computed for each subtest by adding the total number of correct responses. All 244 participants had an opportunity to provide responses to a minimum of items 1 through 3 on each subtest. Because the focus of the current study involved item level analyses, only items 1-3 of each subtest were included, thus assuring that all student participants had had an opportunity to provide an item level answer.

Child language profiles.

Defining children's language ability. Child language abilities were defined as a combination of syntax, morphology, vocabulary, and semantics, and were operationalized using the Clinical Evaluation of Language Fundamentals edition four (CELF-4; Semel, Wiig, & Secord, 2003), the Peabody Picture Vocabulary Test III Form A (PPVT; Dunn & Dunn, 1997), and the Expressive Vocabulary Test (EVT; Williams, 1997). These measures of language are considered below.

Syntactic and morphological functioning. Syntactic and morphological functioning can be conceptually defined as awareness of grammaticality. The CELF-4 Language Structure Index was used to measure children's syntactic and morphological functioning. The CELF-4 is a commonly used measure of language functioning with high construct validity across typical and atypical language users (including gifted students, students with hearing impairments, visual impairments, developmental delays, intellectual disabilities, and autistic disorder; Semel, Wiig,

& Secord, 2003). The CELF-4 standardization sample included more than 4,500 participants (ages 5 to 21 years) from geographically diverse regions in the United States. For this sample of children, with average language age 4.80 years (SD = 1.63), subtests appropriate for children ages 5 to 8 years were selected to indicate receptive language, expressive language, language content (semantics), and language structure (syntax and morphology) of interest on the CELF-4 (appropriate subtests included Concepts and Following Directions, Word Structure, Recalling Sentences, Formulated Sentences, Word Classes I Receptive Vocabulary, Word Classes I Expressive Vocabulary, and Sentence Structure).

The Word Structure subtest, Recalling Sentences subtest, Formulated Sentences subtest, and Sentence Structure subtest comprise the CELF-4 Language Structure Index score. All of the subtests included in the Language Structure Index were administered and used as indicators of the syntactic and morphological aspects of child language profile.

The Word Structure subtest presented students with verbal statements to be completed using the aid of illustrations. Administrators asked the students using verbal statements about one picture, and students responded with grammatically equivalent statements about another picture in the array (e.g., “This boy is walking, and this boy ___” would entail answering with the grammatically equivalent statement “is running”). All 32 items in the subtest were administered in this untimed assessment. Raw scores were computed from totaling correct responses.

The Recalling Sentences subtest presented students with verbal statements to be repeated back to the examiner verbatim. The statements became more grammatically complex, longer, and included more parts of speech as the assessment progressed. Items were scored based on the number of errors made in sentence repetition. Items were scored ‘3’ if no errors were made, ‘2’ if one error was made, ‘1’ if two to three errors were made, and ‘0’ if four or more errors were

made. The assessment was untimed. Ceiling was reached when students answered five consecutive items with four or more errors in repetition. Raw scores were computed by totaling the scores for each item.

The Formulated Sentences subtest presented the students with an illustration and a single word verbal prompt. The single word was to be used in a complete sentence relating to the illustration presented (e.g., “Make a sentence about this picture using the word ‘book.’”). Responses were scored ‘2’ if no grammatical errors were made and the target word was used, ‘1’ if a grammatical error was made and the target word was used, and ‘0’ if two or more grammatical errors were made and/or if the target word was not used. The subtest was untimed and administered until a ceiling of five, consecutive scores of ‘0’ were obtained. Raw scores were computed by totaling the scores for each item.

The Sentence Structure subtest was administered with a visual array of four, similar scenes and an orally presented stimulus. The stimulus was a complete sentence describing one of the scenes depicted, and students responded by selecting the scene described by the verbal prompt. The items varied in grammatical content and difficulty. The subtest was untimed, and all 26 items were administered. Raw scores were computed by totaling the number of correct responses.

In general, reliability for the CELF is dependent on subtest and age of examinee, and so reliability was considered relative to the ages and language skills of the participants in the current study. Selected subtests demonstrate high reliability (.70 and higher internal consistency coefficient alpha) across content, time, and scorer (Semel, Wiig, & Secord, 2003). The Sentence Structure subtest in particular was noted as a way of discriminating between children with and without language disorders. For children identified as having intellectual disabilities, the

Language Structure subtests all displayed reliabilities at and above .85 (Semel, Wiig, & Secord, 2003).

Vocabulary knowledge. Vocabulary knowledge, with regard to both receptive and expressive vocabulary, can be conceptually defined as a combination of both stored phonological and semantic representations of words (Levelt, Roelofs, & Meyer, 1999). The Peabody Picture Vocabulary Test III Form A (PPVT; Dunn & Dunn, 1997) was used to assess receptive vocabulary, and the Expressive Vocabulary Test (EVT; Williams, 1997) was used to assess expressive vocabulary because they are commonly accepted measures of the constructs and also have demonstrated validity across examinees with both typical and atypical language profiles, including individuals with mild intellectual disabilities. The PPVT III and EVT are both appropriate for a broad range of ages (two years and six months through adulthood). These assessments were administered such that basal scores and ceilings were established for all participants.

The PPVT III was administered by presenting students with an array of four illustrations. Students were asked to point to the picture that depicted the target vocabulary item (e.g., “Point to the picture that shows ‘baby’.”). Items were divided into 17 sets with 12 items each. The PPVT III is not a timed assessment. Items were administered until students reached a ceiling of eight incorrect items in a set. Raw scores were calculated by subtracting the total number of incorrectly answered items from the last item in the ceiling set. Higher raw scores indicated higher receptive vocabulary.

For all applicable ages, the reliability for the PPVT III is high across content, time, and scorer. Split half reliability coefficients across ages are all at or above .91 (Dunn & Dunn, 1997). Items on the PPVT III display high internal validity in terms of homogeneity and age

differentiation. The PPVT III correlates well with other measures of vocabulary and moderately well with measures of verbal ability, indicating high construct validity (Dunn & Dunn, 1997).

The first section of the EVT (designed for children ages two years and six months to four years and eleven months) was administered by presenting students with an illustration and asking them to name objects or actions (e.g., “What is this,” or “Tell me a word for this?”). The second section of the EVT (designed for children age five to adults) was administered by presenting an illustration and a verbal label of that illustration. Examinees were then asked to provide another word for the illustration (e.g., “I am going to say a word and I want you to tell me another word that means the same thing. Bag. Tell me another word for ‘bag’.”). The assessment was not timed. Items were administered until students reached a ceiling of five consecutive incorrect responses. Raw scores were calculated by subtracting the total incorrectly answered items from the last item administered. Higher raw scores indicated higher expressive vocabulary.

The EVT demonstrates high reliability in both test-retest results and item uniformity in the normative sample. The EVT also demonstrates high construct validity as evidenced by word frequency data, age differentiation, and correlation with other language measures requiring expression.

Semantic knowledge. Semantic knowledge can be conceptually defined as awareness of meaning at the word, sentence, and connected text levels (Semel, Wiig, & Secord, 2003). The CELF-4 Language Content Index was used to measure children’s semantic knowledge. For typically developing children ages 5 to 8 years (and for children with similar language development), the Concepts and Following Directions subtest, the Word Classes I subtest, and the Expressive Vocabulary subtest comprise the CELF-4 Language Content Index score. However, due to the inclusion of the EVT as a measure of expressive vocabulary knowledge and

considerations of total testing time and child fatigue, the CELF-4 Expressive Vocabulary subtest was not included in the total testing battery for this study. Instead, Concepts and Following Directions and Word Choices I were selected as the subtests to be included as indicators of the semantic aspects of child language profile.

The Concepts and Following Directions subtest presented students with verbal directions of increasing complexity and length to be completed using the aid of illustrations. Administrators asked the students to point to illustrations with specific names and attributes in the order specified by the directions, and students responded by pointing to picture(s) in the illustrated array (e.g., “Point to the pictures that are red,” would entail pointing to only the red items in an array). All 23 of the set 1 items in the subtest were administered, and the 31 items in set 2 were administered to a ceiling of seven consecutive incorrect items. The Concepts and Following Directions subtest was untimed. Raw scores were computed from totaling correct responses.

The Word Choices I subtest presented students with illustrated arrays of objects, a verbal prompt to identify the two objects that “go together,” and a verbal prompt to identify how the two selected objects “go together.” First, administrators labeled objects in the array and asked the students to identify the two objects that "go together." Students responded with either verbal statements or by pointing to identify objects (e.g., “Here are sandwich, apple, and plate. Which two go together?” would entail answering with “sandwich and apple”), completing the Receptive portion of the Word Classes subtest. Next, administrators prompted the students to explain how their selections “go together,” (e.g., “How do sandwich and apple go together?”). Students then completed the Expressive portion of the Word Classes subtest by explaining their rationale for selecting two items as similar, (e.g., “Sandwich and apple go together because they are both

types of food.”). All 21 items in the subtest were administered in this untimed assessment. Raw scores were computed from totaling correct responses.

Selected subtests demonstrate high reliability (.84 and higher internal consistency coefficient alpha) across content, time, and scorer (Semel, Wiig, & Secord, 2003). For children identified as having intellectual disabilities, the Concepts and Following Directions subtest and the Word Choices I subtest both displayed reliabilities at and above .85 (Semel, Wiig, & Secord, 2003).

Linguistic complexity of mathematics test items.

Defining linguistic complexity. For the purposes of this study, linguistic complexity was considered under the relative linguistic complexity definition. Relative linguistic complexity considers not only the form and meaning of language, but also the language abilities of the user. Because this study was evaluating linguistic complexity as it interacted with the language profiles of atypical language users, relative linguistic complexity was most appropriate. Linguistic complexity was defined as the relative complexity of a unit of language with regard to both structure and content (a combination of vocabulary, utterance length, and grammatical complexity). The vocabulary level of linguistic complexity was measured using the relative frequency of occurrence of words in the English language. Utterance length was measured using the number of morphemes, words, and different words occurring in each item. Grammatical complexity was measured using the Developmental Sentence Score of each item. These operational definitions are explored in more detail below.

Word frequency. The Educator’s Word Frequency Guide (Zeno, Ivens, Millard, & Duvvuri, 1995) is a collection of various frequency measures of words in the English language, estimated using over 17,000,000 tokens representing a variety of English text samples across a

variety of disciplines and content areas (e.g., language arts, social sciences, science, mathematics, fine arts, health, safety, etc.). The “D statistic” of word frequency describes the relative entropy of a word, its frequency of use from zero (words appearing rarely and in only one content area) to one (words appearing frequently and across all content areas). For example, the word “the” has a D statistic of .9971, reflecting a relatively large frequency of use, while the word “anorexia” has a D statistic of .2221, reflecting a relatively small frequency of use. The standardized item prompts for each target KM-R item in the analysis were transcribed according to CHAT transcription program conventions (MacWhinney, 2000). D statistics for each word were identified using the Word Frequency Guide (Zeno et al., 1995).

Utterance length. CHAT transcripts for each target KM-R item in the analysis were analyzed for standard utterance length measures using the CLAN software program (MacWhinney, 2000). Number of different words, number of total words, mean length utterance in words, and mean length utterance in morphemes were of particular interest in defining utterance length.

Developmental sentence score. The CLAN Developmental Sentence Score (DSS) program (MacWhinney, 2000) computes the DSS statistic first described by Lee (1974). Sentences are scored based on morphosyntactic and lexical considerations across eight grammatical domains (indefinite pronouns, personal pronouns, main verbs, secondary verbs, negatives, conjunctions, interrogative reversals, and Wh-questions). Higher scores are associated with higher grammatical complexity. KM-R item CHAT transcripts were first analyzed for morphological codes using the CLAN MOR program for English (MacWhinney, 2000). Once a CLAN %mor morphological tier with parts of speech had been created for the transcripts, the CLAN POST program was run to resolve syntactic ambiguities, creating a %post syntactic tier

with disambiguated grammatical markers for each word in an utterance. The final output was reviewed for correct morphosyntactic coding. Finally, the CLAN DSS program was run to identify the Developmental Sentence Scores for each utterance (in the interactive mode) to determine the CLAN DSS score.

KM-R items were evaluated using the CHAT transcription conventions and the CLAN software program. CLAN outputs of number of different words, number of total words, mean length utterance in words, mean length utterance in morphemes, and DSS were each incorporated as item level measures in an item level database for item linguistic complexity analyses.

Design

Students in the parent project were randomly assigned to one of three interventions, all of which included some direct instruction with an emphasis on repetition, review, and teaching to skill mastery. The Phonological Decoding and Blending reading intervention (PHAB; Engelmann & Bruner, 1988a & 1988b) featured explicit teaching of phonemes and blending exercises at the phoneme, word, and connected text levels. The PHAB + RAVEO reading intervention (abbreviated as the RAVEO condition, meaning Retrieval-Rate, Automaticity, Vocabulary Elaboration, and Orthography; Wolf, Miller, & Donnelly 2000) featured the phonological components of the PHAB intervention but also incorporated vocabulary development, fluency, and comprehension skill development. The Mathematics intervention contrast condition (MATH; Engelmann & Carnine, 1992) incorporated the same instructional format as the reading interventions, but with focused content in the areas of numeration, addition, subtraction, multiplication, division, mental math, oral arithmetic, word problems, and connecting math concepts.

Only children who completed the scheduled 120 hours of intervention were included in this research study. Although the parent project featured assessments of language and mathematics at 0 hours of instruction, 60 hours of instruction, 120 hours of instruction, and 12 months after instruction had completed, the current research focused on data only from the baseline and 120 instructional hour time points of assessment.

Data Collection

After obtaining child assent for testing, a battery of standardized and experimental assessments was administered individually with trained graduate students or psychometrists in the school setting in private areas. The same measurement battery was administered at both the baseline and the 120 instructional hour time points. All test administrators received ongoing training in assessment and feedback on assessment performance. Academic measures for the parent study (e.g., mathematics and reading assessments) were administered before language measures at each time point of assessment in order to avoid potential confounds of continuing academic instruction over the school year. For the purposes of this study, the KM-R was administered to students before the PPVT, EVT, and CELF. Administration of the entire testing battery for the parent study (of which these assessments are only a subset) was estimated to require approximately two hours of a student's time.

After assessment data were obtained, data were scored and checked by two separate research personnel. Both raw and standard scores were entered into a secure SPSS database. Two separate data entries with two separate research personnel were performed, and all data entries were crosschecked for accuracy.

CHAPTER 3: RESULTS

Analysis Overview

The central research questions of this study sought to examine (1) the contribution of item *linguistic complexity*, (2) the contribution of *child language skills*, and (3) the potential interaction between *item linguistic complexity* and *child language skill* in predicting item level mathematics assessment performance on the KeyMath-Revised Diagnostic Inventory of Essential Mathematics (KM-R). The longitudinal stability of these relationships was specified as a planned post hoc analysis. Prior to analyzing the multilevel research model, relevant child-level and item-level covariates were identified in a series of correlation analyses. Next, child language profile and item linguistic complexity were analyzed for confirmation of specified factor structures. Next, because children's language profiles could change over time and intervention experience, measurement invariance of the child language profile factor was analyzed in a single sample longitudinal invariance model prior to analyses for structural invariance. Finally, the relationship between *item linguistic complexity* and *child language profile* was modeled in a multilevel interaction presented in Figure 1. The proposed structural model was examined at both baseline and post intervention, and the specified structural relationship at these two time points was also examined for structural invariance in order to test the stability of item linguistic complexity and child language profile as predictors of mathematics achievement over time and educational experiences. Each of these analyses are presented in the sections that follow.

Covariate Analyses

Group and School Level Covariates.

Although the primary interests of this study were at the item-level (within) and child-level (between) the higher between-levels of school and intervention group also were examined

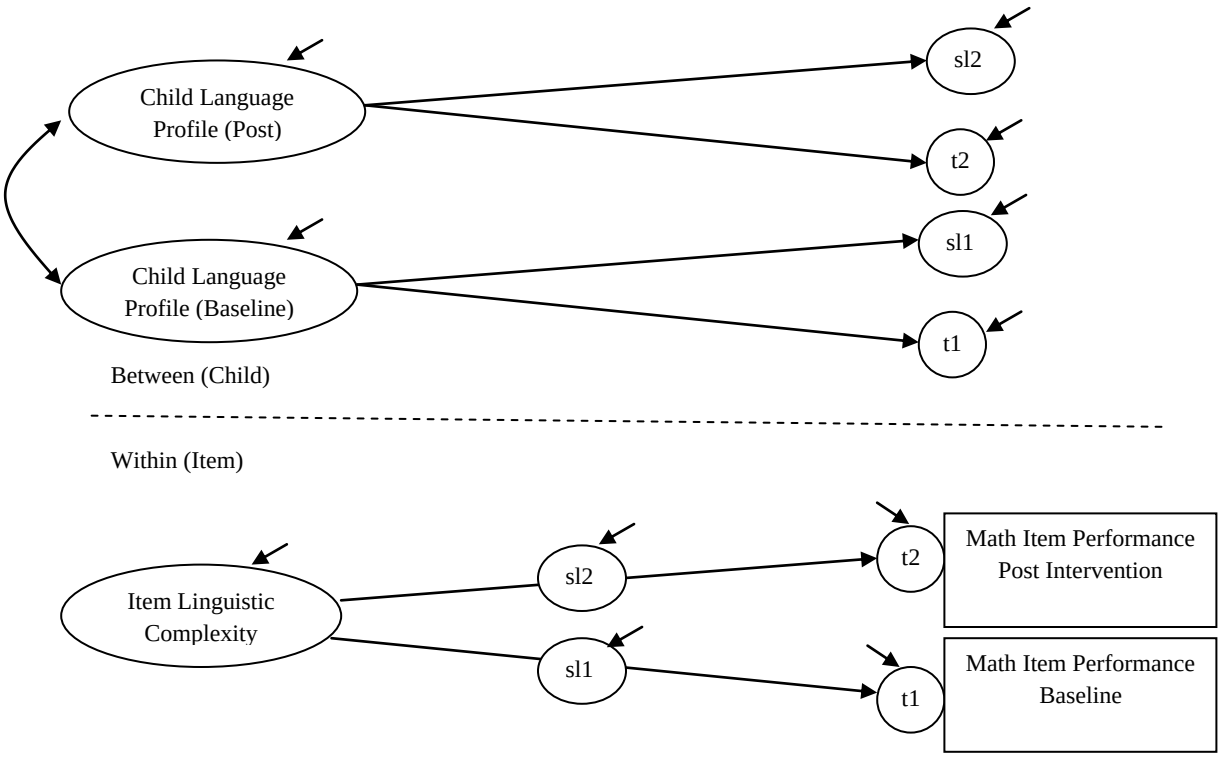

Figure 1. Proposed two-level model of mathematics item performance.

for disaggregated contributions at the child-level. Intervention group and school differences were examined with particular interest in identifying child-level covariates. Intervention group composition analyses are included in Appendix B. The results of school characterization analyses are presented in Appendix C. The intervention group analyses indicated some significant group differences in age, grade level, mother education, and father education. The school analyses indicated some significant differences in intervention groups in the following school-level variables: Title I status, percentage of students eligible for free or reduced lunch, percentage of White students, percentage of Hispanic students, and percentage of Multiracial

students. Special attention was paid to student age, grade level, socioeconomic status, and race in subsequent child-level covariate analyses.

Child Level Correlations and Covariates.

Potential covariates were examined in a series of bivariate correlation analyses. First, potential covariates were examined with planned indicators of the *child language profile* latent factor (CELF-4 raw scores for the Concepts and Following Directions, Word Structure, Recalling Sentences, Formulating Sentences, Word Classes I, and Sentence Structure subtests, and total raw scores for the PPVT III and EVT). Significant correlations greater than .33 were considered to be of interest for inclusion as covariates. The following child demographic variables were considered for inclusion as covariates: age, sex, race, IQ, grade level, and socioeconomic variables measured by the Hollingshead index. Of these child demographic variables, only age, grade level and IQ met the criteria for additional examination as covariates in both the baseline *child language profile* and post *child language profile* correlation analyses (see Tables 3 and 4 for full child-language by child-demographic correlation results). Age correlated at between .33 and .47 with *child language profile* indicators at baseline and at between .24 and .41 with *child language profile* indicators post intervention. Grade level correlated at between .32 and .45 with *child language profile* indicators at baseline and at between .21 and .46 with *child language profile* indicators post intervention. Finally, IQ correlated at between .25 and .46 with *child language profile* indicators at baseline and at between .21 and .46 with *child language profile indicators* post intervention.

Next, potential covariates were examined with the outcome variable of interest, item-level child responses to KeyMath questions (at both baseline and post intervention). Full bivariate results are displayed in Tables 5 and 6. At baseline, only age, grade level, and IQ

correlated with the KM-R items of interest above .33. Both age and grade level correlated with Subtraction Item 1 at $r = .41$, $p < .01$. Student IQ correlated with Subtraction Item 1 at $r = .37$, p

Table 3

Child Language Indicators (Baseline) By Child Demographic Variables Correlation Matrix

		Baseline Child Language Profile Indicators (CELF Subtest Scores, PPVT Raw Scores, and EVT Raw Scores)							
		Concepts & Following Directions	Word Structure	Recalling Sentences	Formulating Sentences	Word Classes	Sentence Structure	PPVT	EVT
Child Demographic Covariates Of Interest	Baseline Age (Mon.)	.42**	.34**	.33**	.43**	.41**	.41**	.47**	.44**
	Sex	.09	.10	.17*	.12	.11	.09	.06	.05
	African Am. Race	.03	-.02	.16*	.05	.04	.11	.06	.03
	Caucasian Race	.02	.14*	-.01	.06	-.05	-.02	.15*	<.01
	Hispanic Race	-.06	-.15*	-.15*	-.15*	.03	-.09	-.23**	-.07
	Asian Race	-.03	-.13	-.10	-.04	-.05	-.06	-.10	-.07
	Mixed Race	.01	.12	-.01	.07	-.03	<.01	.06	.11
	Student IQ	.37**	.32**	.25**	.30**	.46**	.32**	.28**	.34**
	Grade Level	.39**	.33**	.32**	.42**	.41**	.41**	.45**	.44**
	Mother Ed. (Yrs)	-.04	<.01	-.01	.07	-.09	-.03	.07	.02
	Moth. HH Occ. Score	.01	.04	.03	.08	<.01	.06	.11	.05
	Fath. Ed. (Yrs)	-.09	.03	.02	.07	-.10	-.09	.12	.03
	Fath. HH Occ. Score	-.13	.02	-.09	.02	-.12	-.04	.11	.01
	Moth. Overall HH Score	-.02	.02	.01	.05	-.04	.03	.08	.03
	Fath. Overall HH Score	-.15	.02	-.07	.03	-.14	-.08	.11	<.01
	Fam. Overall HH Score	-.05	.02	-.03	.02	-.09	-.01	.11	.02

Note. **. Correlation is significant at the .01 level (2-tailed).
*. Correlation is significant at the .05 level (2-tailed).
CELF: Clinical Evaluation of Language Fundamentals 4th edition (Semel, Wiig, & Secord, 2003).
PPVT: Peabody Picture Vocabulary Test III Form A (Dunn & Dunn, 1997).
EVT: Expressive Vocabulary Test (Williams, 1997).
HH: Hollingshead Two Factor Index of Social Position (Hollingshead, 1975).

Table 4

Child Language Indicators (Post Intervention) By Child Demographic Variables Correlation Matrix

		Post Intervention Child Language Profile Indicators (CELF Subtest Scores, PPVT Raw Scores, and EVT Raw Scores)							
		Concepts & Following Directions	Word Structure	Recalling Sentences	Formulating Sentences	Word Classes	Sentence Structure	PPVT	EVT
Child Demographic Covariates Of Interest	Baseline Age (Mon.)	.33**	.31**	.24*	.28**	.24*	.25**	.41**	.38**
	Sex	.06	.08	.06	-.004	.08	.12	.04	.07
	African Am. Race	-.06	-.05	.05	-.10	.01	.001	-.10	-.07
	Caucasian Race	.07	.10	-.07	.07	-.02	-.06	.11	.05
	Hispanic Race	.03	-.13	-.01	.11	.08	.14	-.05	.04
	Asian Race	-.03	.01	-.15	-.04	<.01	-.08	-.03	-.01
	Mixed Race	-.04	.08	.14	-.05	-.08	-.02	.08	-.003
	Student IQ	.46**	.34**	.31**	.34**	.37**	.41**	.32**	.32**
	Grade Level	.26**	.26**	.21*	.28**	.25**	.22*	.46**	.38**
	Mother Ed. (Yrs)	-.14	.08	-.07	-.14	-.10	-.22*	.03	.03
	Moth. HH Occ. Score	-.04	.07	.02	.01	.02	-.02	.06	.09
	Fath. Ed. (Yrs)	-.19*	.02	-.09	-.18	-.14	-.28**	-.10	-.11
	Fath. HH Occ. Score	-.11	-.01	-.13	-.06	-.12	-.23*	.01	.03
	Moth. Overall HH Score	-.07	.08	-.01	-.03	-.01	-.07	.05	.08
	Fath. Overall HH Score	-.16	-.004	-.14	-.11	-.14	-.27**	-.03	-.01
	Fam. Overall HH Score	-.13	.05	-.08	-.08	-.08	-.20*	.01	.05

Note. **. Correlation is significant at the .01 level (2-tailed).
*. Correlation is significant at the .05 level (2-tailed).
CELF: Clinical Evaluation of Language Fundamentals 4th edition (Semel, Wiig, & Secord, 2003).
PPVT: Peabody Picture Vocabulary Test III Form A (Dunn & Dunn, 1997).
EVT: Expressive Vocabulary Test (Williams, 1997).
HH: Hollingshead Two Factor Index of Social Position (Hollingshead, 1975).

Table 5

Child Item Response (at Baseline) By Child Demographic Correlation Matrix

		Baseline Item Response Variables																	
		Nm 1	Nm 2	Nm 3	Ge 1	Ge 2	Ge 3	Ad 1	Ad 2	Ad 3	Sb 1	Sb 2	Sb 3	Me 1	Me 2	Me 3	TM 1	TM 2	TM 3
Child Demographic Covariates of Interest	Baseline Age (Mon.)	.15	.18	.11	.11	.19	.14	.21*	.21*	.29**	.41**	.02	.18	.12	.18	.23*	.19	.13	.17
	Sex	-.04	-.04	.06	.13	.07	.06	-.04	-.08	.05	-.01	<.01	-.05	-.06	-.12	.07	.06	.04	.22*
	African Am. Race	-.21*	-.19*	.09	.17	.03	.07	.06	-.04	-.03	-.07	.18	-.17	-.10	-.05	-.19*	.14	-.13	-.08
	Caucasian Race	.14	.06	-.13	-.03	<.01	-.17	.06	-.03	.03	.01	-.12	.22*	-.02	.05	.12	-.02	.10	-.03
	Hispanic Race	.07	.16	.04	-.11	-.09	.07	-.23*	-.06	.03	.12	-.06	-.10	.12	-.02	.12	-.27**	.03	.14
	Asian Race	.03	-.15	.06	-.12	<-.01	-.01	.10	.13	-.07	.04	-.03	-.04	.02	.04	-.01	-.04	.13	-.01
	Mixed Race	.05	.16	-.03	-.06	.05	.10	-.01	.13	.01	-.06	-.04	.08	.06	<.01	.01	.14	-.04	.04
	Student IQ	.05	.24*	.23*	.18	.16	.30**	.07	.15	.29**	.37**	-.03	.12	.41**	.25**	.18	.11	.20*	.37**
	Grade Level	.14	.18	.23*	.05	.17	.21*	.20*	.20*	.17	.41**	.01	.14	.16	.14	.15	.09	.21*	.15
	Mother Ed. (Yrs)	.10	-.07	.04	-.05	.07	-.16	.10	-.10	-.07	-.15	-.11	.10	-.26**	-.12	-.19	.12	.01	-.01
	Moth. HH Occ. Score	.10	.08	.02	.03	.11	-.03	.13	.08	-.15	-.07	-.08	-.01	-.08	-.02	-.04	.10	.13	.07
	Fath. Ed. (Yrs)	.01	-.19	.06	.04	-.08	-.20*	.23*	.02	-.13	-.21*	-.06	<.01	-.31**	-.05	-.16	.13	-.01	-.09
	Fath. HH Occ. Score	.13	-.05	.04	-.12	-.20*	-.12	.08	.03	-.03	-.17	-.08	.02	-.16	-.11	-.07	<-.01	.01	-.05
	Moth. Overall HH Score	.11	.04	.02	.01	.10	-.07	.14	.04	-.14	-.10	-.10	.02	-.14	-.06	-.09	.11	.10	.06
	Fath. Overall HH Score	.11	-.10	.05	-.08	-.18	-.16	.14	.03	-.05	-.22*	-.09	.01	-.22*	-.10	-.10	.04	.01	-.06
	Fam. Overall HH Score	.13	-.03	.04	-.04	-.04	-.12	.17	.04	-.11	-.19	-.11	.02	-.21*	-.09	-.11	.09	.07	<.01

Note. **. Correlation is significant at the .01 level (2-tailed).
*. Correlation is significant at the .05 level (2-tailed).
HH: Hollingshead Two Factor Index of Social Position (Hollingshead, 1975).

Table 6

Child Item Response (Post-Intervention) By Child Demographic Correlation Matrix

		Post Intervention Item Response Variables																	
		Nm 1	Nm 2	Nm 3	Ge 1	Ge 2	Ge 3	Ad 1	Ad 2	Ad 3	Sb 1	Sb 2	Sb 3	Me 1	Me 2	Me 3	TM 1	TM 2	TM 3
Child Demographic Covariates of Interest	Baseline Age (Mon.)	.08	.09	.15*	.17*	.23**	.17*	.13	.23**	.22**	.26**	.13*	.17*	.23**	.23**	.11	.10	.18*	.22**
	Sex	-.05	-.02	.05	.09	<.01	-.03	-.01	.10	-.01	-.05	.01	.01	.07	-.02	.04	.04	-.07	.06
	African Am. Race	-.06	-.09	-.03	.05	-.03	-.06	-.02	-.12	-.01	-.13	<.01	-.05	.01	-.04	-.03	-.02	-.04	.02
	Caucasian Race	.03	.08	.02	-.01	.02	<.01	<.01	.06	-.05	.04	.06	.11	-.02	.04	.01	.03	-.01	-.03
	Hispanic Race	.03	-.03	.04	-.05	.03	.06	.03	.05	.05	.05	-.03	-.09	<.01	-.01	.04	-.06	.06	.04
	Asian Race	.01	.06	.03	-.01	.07	.07	.05	.06	.03	.08	-.03	-.03	.03	.02	.02	.02	.09	-.10
	Mixed Race	.01	.08	-.06	-.01	-.06	-.01	-.05	.02	<.01	.07	-.04	.07	-.01	.03	-.02	.11	-.05	.02
	Student IQ	.04	.27**	.15*	.15*	.34**	.26**	.29**	.23**	.19*	.28**	.14*	.13	.27**	.24**	.24**	.07	.14	.15*
	Grade Level	.08	.06	.15*	.17*	.21**	.16*	.11	.20**	.24**	.26**	.14*	.15*	.24**	.24**	.06	.11	.21**	.22**
	Mother Ed. (Yrs)	.02	-.11	-.04	-.02	-.09	-.07	-.07	-.04	-.10	-.10	.05	.08	-.16*	-.13	-.14*	.11	.01	<.01
	Moth. HH Occ. Score	.06	-.12	-.06	-.13	-.05	-.08	-.10	.04	-.06	.02	-.12	.06	-.15*	-.11	.01	.09	.11	<.01
	Fath. Ed. (Yrs)	.01	-.01	.03	-.03	-.11	-.05	-.03	.02	-.13	-.02	.05	.02	-.16	-.09	-.06	.13	-.05	-.05
	Fath. HH Occ. Score	.12	-.06	.06	-.09	-.08	-.13	.01	.08	.02	-.01	-.02	.03	-.04	-.05	-.10	.12	-.03	-.11
	Moth. Overall HH Score	.06	-.14*	-.06	-.13	-.07	-.09	-.12	.01	-.06	-.02	-.10	.07	-.18*	-.16*	-.04	.09	.09	-.01
	Fath. Overall HH Score	.10	-.07	.06	-.10	-.08	-.12	-.02	.06	-.02	-.02	-.01	.03	-.07	-.07	-.09	.14	-.02	-.11
	Fam. Overall HH Score	.07	-.12	-.03	-.16*	-.10	-.13	-.08	.02	-.04	-.02	-.05	.08	-.15*	-.15*	-.05	.11	.06	-.06

Note. **. Correlation is significant at the .01 level (2-tailed).
*. Correlation is significant at the .05 level (2-tailed).

HH: Hollingshead Two Factor Index of Social Position (Hollingshead, 1975)

< .01, Measurement Item 1 at r = .41, p < .01, and Time and Money Item 3 at r = .37, p < .01. At the post intervention time point, the correlation between student IQ and Geometry Item 2 response was the only correlation of magnitude above .33, r = .34, p < .01.

Finally, potential child-level demographic covariates were examined for redundancy in a bivariate correlation matrix with themselves. The primary foci of this analysis were age, grade level and IQ, as these had been implicated in correlations with child-language profile variables and the outcome variables of interest. As might be expected, the bivariate correlation between child age and grade level indicated colinerarity between these two variables, r=.86, p<.01; therefore, only child grade-level was controlled for as a covariate in subsequent multilevel analyses.

IQ did not have any significant correlations above .30 with other child-level demographic variables of interest. However, because (1) IQ only displayed low to moderate correlations with predictor and outcome variables in this analysis, (2) only 206 of the 244 children in the sample had IQ data, and (3) IQ information was not missing at random (schools displayed different patterns of IQ missing data and some schools did not provide IQ data for any of their child participants in the study), it was not selected as a control variable in the final, multilevel model analysis. The contribution of student IQ was more closely examined in factor analyses of *child language profile*, in which the *child language profile* latent factor was examined both with and without IQ as a control variable. Table 7 displays the full child-demographic variable bivariate correlation matrix.

Item Level Correlations and Covariates.

Item linguistic complexity was operationally defined with the following indicators: mean length utterance (in words and morphemes), number of total words, developmental sentence

Table 7

Child Demographic Variable Correlation Matrix

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Baseline Age (Mon.)	.03	- .05	.08	-.04	-.08	.09	-.07	.86**	.03	.05	.02	.05	-.01	.05	.08
2. Sex		.12	-.08	-.12	.11	-.01	.04	.01	.06	.04	-.05	-.10	.05	-.10	.02
3. African Am. Race			-.59**	-.50**	-.16*	-.25**	-.10	.04	.22**	.12	.16*	-.25**	.16*	-.16	.07
4. Caucasian Race				-.23**	-.08	-.11	.01	-.02	.11	.01	.17*	.40**	.03	.37**	.18**
5. Hispanic Race					-.06	-.10	.21**	-.08	-.52**	-.26**	-.54**	-.22**	-.34**	-.35**	-.40**
6. Asian Race						-.03	-.02	-.04	.03	.05	.06	.04	.03	.04	.06
7. Mixed Race							-.12	.10	.16*	.09	.15	.08	.11	.11	.12
8. Student IQ								-.06	-.19*	-.14	-.29**	-.20*	-.18*	-.27**	-.21**
9. Grade Level									.10	.02	-.04	.04	.04	.02	.04
10. Mother Ed. (Yrs)										.51**	.74**	.38**	.69**	.54**	.70**
11. Moth. HH Occ. Score											.39**	.32**	.97**	.39**	.86**
12. Fath. Ed. (Yrs)												.48**	.50**	.69**	.68**
13. Fath. HH Occ. Score													.37**	.96**	.77**
14. Moth. Overall HH Score														.47**	.91**
15. Fath. Overall HH Score															.85**
16. Fam. Overall HH Score															

Note. **. Correlation is significant at the .01 level (2-tailed).
*. Correlation is significant at the .05 level (2-tailed).
HH: Hollingshead Two Factor Index of Social Position (Hollingshead, 1975).

score, and minimum word frequency. A number of additional item-level language characteristics were investigated for potential correlation with the indicators of *item linguistic complexity* as a latent factor. Of particular interest in these bivariate correlation analyses were total number of utterances, number of different words, type token ratio, and word frequency mean. A table of item level descriptive indices is presented in Table 8. A full table of bivariate correlations for item level indicators of *linguistic complexity* is displayed in Table 9.

Although the total number of utterances displayed a moderate to high correlation with number of total words, $r = .62$, $p < .05$, it was not used as a covariate in subsequent analyses because of the relatively low variance in total utterances across the Key Math Inventory items examined (all items were 1-2 utterances in length). Number of different words appeared to be collinear with MLU in morphemes, $r = .95$, $p < .01$, and displayed the same pattern of correlations with other variables examined in this analysis; thus it was redundant as a potential covariate with an indicator already established in the examination of *item linguistic complexit*y as a latent factor.

The type token ratio correlated moderately with number of total words, $r = -.57$, $p < .05$, and did not significantly correlate with any of the other selected indicators of *item linguistic complexity*. However, as the type token ratio is defined as the ratio of the number of different words to the number of total words in an item, it also can be considered redundant with the MLU in morphemes and the number of total words already selected as indicators of *item linguistic complexity*.

Finally, the word frequency mean correlated moderately with word frequency minimum (the relative frequency of the least often occurring word in each item), $r = .65$, $p < .01$; however, across the KM-R items under investigation, the word frequency mean was consistently high and

Table 8

KM-R Item Linguistic Complexity Indices

Item	Word Freq. D Min.	No. Total Words	MLU Words	MLU Morphemes	DSS	Percent Answering Correctly
Numeration 1	.62	7	7	7	3	99.6
Numeration 2	.66	12	12	13	8	87.3
Numeration 3	.73	5	5	6	6	95.5
Geometry 1	.56	12	6	6.5	5	81.6
Geometry 2	.65	12	12	13	15	54.5
Geometry 3	.82	7	7	10	6	82.4
Addition 1	.75	19	9.5	20	14	88.1
Addition 2	.75	12	12	15	10	86.1
Addition 3	.56	15	7.5	18	8	29.9
Subtraction 1	.60	14	7	15	21	75.0
Subtraction 2	.74	14	7	16	18	3.7
Subtraction 3	.72	10	10	10	12	3.7
Measurement 1	.65	14	7	8	6	72.5
Measurement 2	.76	10	5	6	6	73.4
Measurement 3	.72	26	13	15	12	74.2
Time & Money 1	.89	9	9	10	6	54.1
Time & Money 2	.85	6	6	7	5	71.3
Time & Money 3	.76	10	5	5	6	55.3

Note. KM-R: KeyMath-Revised: A diagnostic inventory of essential mathematics (Connolly, 1988).

Table 9

Item Linguistic Complexity Indicators By Item Characteristic Variables Correlation Matrix

		Item Linguistic Complexity Indicators				
		MLU in Words	MLU in Morphemes	Number of Total Words	Developmental Sentence Score	Word Frequency Guide D minimum
Item Potential Covariates	Total Utterances	-0.28	0.23	0.62*	0.29	-0.35
	Number of Utterances	-0.23	-0.42	0.32	-0.29	-0.13
	Number of Different Words	0.59*	0.95**	0.73**	0.77**	-0.3
	Type Token Ratio	0.05	0.14	-0.57*	0.13	0.13
	Word Frequency Guide D maximum	0.19	0.34	0.44	0.36	-0.04
	Word Frequency Guide D median	0.32	0.22	0.27	0.24	0.2
	Word Frequency Guide D mean	0.31	-0.03	-0.05	-0.05	0.65**

Note. **. Correlation is significant at the .01 level (2-tailed).
*. Correlation is significant at the .05 level (2-tailed).

Item indices are reported for the selected items on the KM-R, KeyMath-Revised: A diagnostic inventory of essential mathematics (Connolly, 1988).

displayed little variance, M = .92, SD = .03. Due to the consistent use of common words such as "the," "and," and "to" in the small utterance items of the KM-R, the word frequency mean is less informative than the minimum as a measure of word frequency, and the moderate correlation between the two is likely best explained by the fact that the word with the minimum word frequency in each item contributes to the calculation of that item's word frequency mean. Thus, word frequency mean was not selected as a potential covariate for *item linguistic complexity* in subsequent analyses.

Preliminary Measurement Model Analyses

Prior to examining the larger structural model under investigation in this research study, the measurement models for *item linguistic complexity* and *child language profile* were examined using factor analyses. Because the underlying factor structures of the latent constructs were theoretically defined and predicted using rationales outlined previously, a confirmatory factor analysis was appropriate for investigation of the measurement model. Robust maximum likelihood estimation (MLV) was performed using Mplus (v.6; Muthen & Muthen, 2010).

Child Language Profile at Baseline CFA.

The *child language profile* latent factor (at baseline) was examined in a CFA. Figure 2 displays the proposed child language profile factor structure. This one factor model demonstrated relatively high factor loadings, with all indicators displaying completely standardized loadings between .78 and .84. Indicator variances and correlations were all in the admissible range. Table 10 displays the indicator means, standard deviations, and correlations. Table 11 displays the standardized and unstandardized indicator-factor loadings with confidence intervals. The chi-square exact fit test indicated significant misfit between the 1-factor model and the data, χ^2 (20) = 84.51, $p < .001$. However, approximate fit statistics indicated that the 1-factor model was an

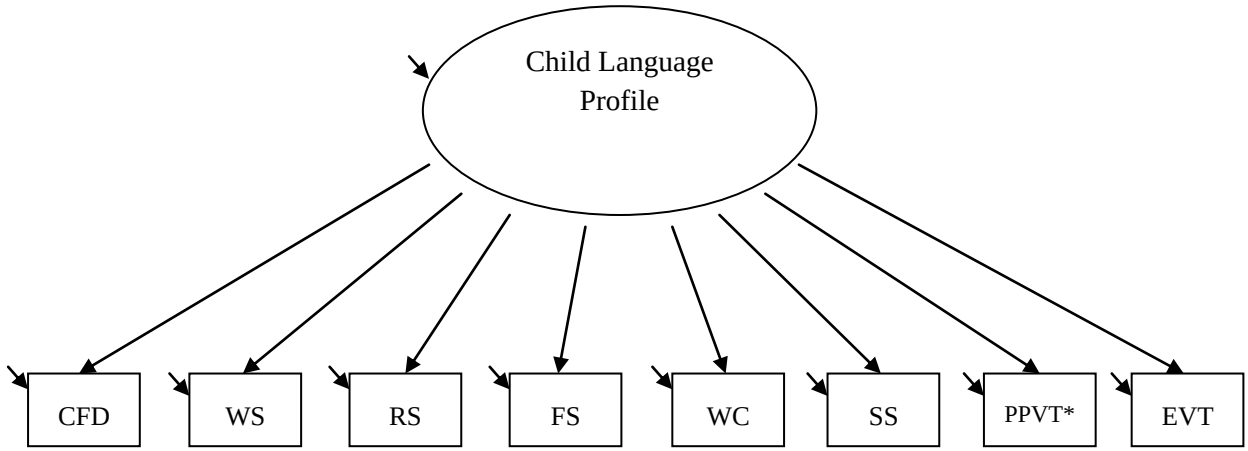

Figure 2. Child language profile proposed factor structure.

Note. CELF: Clinical Evaluation of Language Fundamentals 4th edition (Semel, Wiig, & Secord, 2003).
PPVT: Peabody Picture Vocabulary Test III Form A (Dunn & Dunn, 1997).
EVT: Expressive Vocabulary Test (Williams, 1997).
*Indicates the indicator set as a reference indicator for the scale.
CFD: Concepts and Following Directions subtest of the CELF.
WS: Word Structure subtest of the CELF raw score.
RS: Recalling Sentences subtest of the CELF raw score.
FS: Formulated Sentences subtest of the CELF raw score.
WC: Word Choices I subtest of the CELF raw score.
SS: Sentence Structure subtest of the CELF raw score.

Table 10

Baseline Child Language Profile Indicator Means, SDs, and Correlations

	Mean (SD)	CFD	WS	RS	FS	WC	SS	PPVT	EVT
Concepts & Following Directions	12.79 (9.09)		.69**	.70**	.63**	.67**	.72**	.65**	.65**
Word Structure	10.69 (6.64)			.73**	.70**	.63**	.62**	.67**	.68**
Recalling Sentences	17.61 (14.92)				.71**	.56**	.57**	.52**	.60**
Formulating Sentences	11.48 (10.51)					.61**	.62**	.61**	.65**
Word Choices	20.97 (10.86)						.70**	.64**	.63**
Sentence Structure	14.18 (5.18)							.68**	.64**
PPVT	67.25 (21.96)								.69**
EVT	49.46 (10.46)								

Note. ** Correlations is significant at the $p < .01$ level.
CELF: Clinical Evaluation of Language Fundamentals 4th edition (Semel, Wiig, & Secord, 2003).
PPVT: Peabody Picture Vocabulary Test III Form A (Dunn & Dunn, 1997).
EVT: Expressive Vocabulary Test (Williams, 1997).

Table 11

Baseline Child Language Profile CFA Standardized and Unstandardized Factor Loadings, Standard Errors, and Confidence Intervals

	Unstandardized Factor Loading	Standardized Factor Loading	SE	$C.I._{95}$
Concepts & Following Directions	.44	.84	.02	.80, .88
Word Structure	.32	.84	.02	.80, .88
Recalling Sentences	.67	.78	.03	.72, .84
Formulating Sentences	.49	.80	.03	.74, .86
Word Choices	.49	.78	.02	.74, .82
Sentence Structure	.24	.81	.03	.75, .87
PPVT*	1.00	.79	.03	.73, .85
EVT	.48	.80	.03	.74, .86

Note. χ^2 (20) = 84.51, $p < .001$, *RMSEA* = .11, *CFI* = .95.
*Indicates the indicator set as a reference indicator for the scale.
CELF: Clinical Evaluation of Language Fundamentals 4th edition (Semel, Wiig, & Secord, 2003).
PPVT: Peabody Picture Vocabulary Test III Form A (Dunn & Dunn, 1997).
EVT: Expressive Vocabulary Test (Williams, 1997).

adequate fit for the data, *RMSEA* = .11, *CFI* = .95 (common practice for approximate fit statistics is to judge good fitting models as displaying RMSEA values < .05 and CFI values > .95, Hu & Bentler, 1995).

Several modification indices (three suggested modifications > 10.00) suggested that item disturbances be allowed to covary (specifically, the Recalling Sentences, Formulated Sentences, Word Structure, and Sentence Structure subtests of the CELF-4). However, in the interest of parsimony, none of the suggested modifications were deemed to be theoretically justifiable in a one factor model of *child language profile*. The proposed factor structure for child language profile was accepted without additional modification.

Child Language Profile Post Intervention CFA.

Similarly, the *child language profile* latent factor (post intervention) was examined in a CFA. This one factor model demonstrated relatively high factor loadings, with all indicators displaying completely standardized loadings between .75 and .85. Indicator variances and correlations were all in the admissible range. Table 12 displays indicator means, standard deviations, and correlations. Table 13 displays the standardized and unstandardized indicator-factor loadings with confidence intervals. Figure 2 displays the proposed child language profile factor structure. Again, the chi-square exact fit test indicated significant misfit between the 1-factor model and the data, χ^2 (20) = 64.11, p < .001. However, approximate fit statistics indicated that the 1-factor model was an adequate fit for the data, *RMSEA* = .10, *CFI* = .97. The pattern of results in child language profile CFAs post intervention is much the same as seen at baseline.

Modification indices were examined for the post intervention *child language profile* 1-factor model, and the theoretical underpinnings of suggested modifications were considered. One modification index (> 10.00) suggested that item disturbances between the PPVT and Recalling

Table 12

Post Intervention Child Language Profile Indicator Means, SDs, and Correlations

	Mean (SD)	CFD	WS	RS	FS	WC	SS	PPVT	EVT
Concepts & Following Directions	16.71 (10.42)		.67**	.71**	.67**	.65**	.73**	.65**	.69**
Word Structure	14.76 (7.16)			.67**	.65**	.63**	.64**	.69**	.69**
Recalling Sentences	22.66 (14.88)				.62**	.54**	.64**	.57**	.63**
Formulating Sentences	16.13 (11.62)					.54**	.59**	.56**	.64**
Word Choices	26.61 (10.29)						.71**	.65**	.63**
Sentence Structure	17.07 (5.21)							.72**	.69**
PPVT	73.76 (20.61)								.73**
EVT	57.04 (11.88)								

Note. ** Correlations is significant at the $p < .01$ level.
CELF: Clinical Evaluation of Language Fundamentals 4th edition (Semel, Wiig, & Secord, 2003).
PPVT: Peabody Picture Vocabulary Test III Form A (Dunn & Dunn, 1997).
EVT: Expressive Vocabulary Test (Williams, 1997).

Table 13

Post Intervention Child Language Profile CFA Standardized and Unstandardized Factor Loadings, Standard Errors, and Confidence Intervals

	Unstandardized Factor Loading	Standardized Factor Loading	SE	$C.I._{95}$
Concepts & Following Directions	.53	.85	.02	.81, .89
Word Structure	.35	.82	.02	.78, .86
Recalling Sentences	.69	.77	.03	.71, .83
Formulating Sentences	.52	.75	.03	.69, .81
Word Choices	.47	.77	.03	.71, .83
Sentence Structure	.26	.84	.03	.78, .90
PPVT*	1.00	.82	.03	.76, .88
EVT	.59	.84	.02	.80, .88

Note. χ^2 (20) = 64.11, $p < .001$, *RMSEA* = .10, *CFI* = .97.
*Indicates the indicator set as a reference indicator for the scale.
CELF: Clinical Evaluation of Language Fundamentals 4th edition (Semel, Wiig, & Secord, 2003).
PPVT: Peabody Picture Vocabulary Test III Form A (Dunn & Dunn, 1997).
EVT: Expressive Vocabulary Test (Williams, 1997).

Sentences subtest be allowed to covary. However, in the interest of parsimony and given the one factor measurement model, this modification was deemed to be theoretically unjustifiable. The proposed factor structure for child language profile was accepted without additional modification.

Key Math Item Linguistic Complexity CFA.

Item linguistic complexity was predicted to be indicated by a combination of vocabulary, utterance length, and grammatical complexity. A single factor solution in which all indicators were allowed to load onto the latent item linguistic complexity factor with no allowed covariance between disturbance terms was analyzed. Each indicator was treated as continuous data, and thus maximum likelihood estimation was performed using Mplus (v.6). Table 14 displays item means, standard deviations, and correlations.

The one factor model, representing a single item linguistic complexity latent factor, demonstrated relatively high factor loadings, with all indicators displaying completely standardized loadings between .62 and .95. The one exception was frequency minimum, which displayed a completely standardized loading of -.19, $CI_{.95}$ = -.64, .26. Indicator variances and correlations were all in the admissible range, again with the exception of frequency minimum, which displayed a high residual variance (completely standardized residual variance of .96) and relatively low correlations with other indicators of item linguistic complexity (ranging between -.03 and -.26, none significantly different from zero). Table 15 displays the standardized and unstandardized indicator-factor loadings with confidence intervals.

The chi-square exact fit test indicated good model fit between the 1-factor model and the data, χ^2 (5) = 2.32, p = .80. Approximate fit statistics also indicated that the 1-factor model was a good fit for the data, *RMSEA* < .001, *CFI* = 1.00 (common practice for approximate fit statistics

Table 14

Item Linguistic Complexity Indicator Means, SDs, and Correlations

	Mean (SD)	Freq. Min.	NTW	MLUW	MLUM	DSS
Word Frequency D Min.	.71 (.09)		-0.26	-0.03	-0.17	-0.17
No. Total Words	11.89 (4.98)			.54*	.68**	.51*
MLU Words	8.17 (2.66)				.59*	0.37
MLU Morphemes	11.14 (4.61)					.70**
Developmental Sentence Score	9.28 (5.02)					

Note. **. Correlation is significant at the .01 level (2-tailed).

*. Correlation is significant at the .05 level (2-tailed).

Table 15

Item Linguistic Complexity CFA Standardized and Unstandardized Factor Loadings, Standard Errors, and Confidence Intervals

	Unstandardized Factor Loading	Standardized Factor Loading	SE	$C.I._{95}$
Word Frequency D Min.	-.004	-.19	.23	-.64, .26
No. Total Words	.83	.72	.10	.52, .92
MLU Words	.38	.62	.12	.38, .86
MLU Morphemes*	1.00	.95	.07	.81, 1.09
Developmental Sentence Score	.83	.73	.07	.59, .87

Note. χ^2 (5) = 2.32, p = .80, *RMSEA* < .001, *CFI* = 1.00.

*Indicates the indicator set as a reference indicator for the scale.

is to judge good fitting models as displaying RMSEA values < .05 and CFI values > .95, Hu & Bentler, 1995). No modification indices for this model could significantly improve the model fit with the data. However, the poor factor loading, low indicator intercorrelations, and high residual variance of the frequency minimum indicator were a basis for an additional factor analysis, which excluded frequency minimum as an indicator of item linguistic complexity.

The second factor analysis considered item linguistic complexity as a single latent factor indicated by MLU in words, MLU in morphemes, number of total words, and Developmental Sentence Score. Word frequency minimum was excluded as an indicator. This model demonstrated high factor loadings, with all indicators displaying completely standardized loadings between .62 and .95. Indicator variances and correlations were all in the admissible range. Table 16 displays the standardized and unstandardized indicator-factor loadings with confidence intervals for the modified item linguistic complexity model. Figure 3 displays the final proposed factor structure for the item linguistic complexity latent factor.

The chi-square exact fit test indicated good model fit between the 1-factor model and the data, χ^2 (2) = 1.17, p = .56. Approximate fit statistics also indicated that the 1-factor model was a good fit for the data, *RMSEA* < .001, *CFI* = 1.00 (common practice for approximate fit statistics is to judge good fitting models as displaying RMSEA values < .05 and CFI values > .95, Hu & Bentler, 1995). Again, no modification indices for this model could significantly improve the model fit with the data. Based on the poor factor loading, low indicator intercorrelations, and high residual variance of the frequency minimum indicator, it was determined that with these relatively small utterances, word frequency was not a strong indicator of item linguistic complexity. The modified model was accepted as the measurement model for item linguistic complexity in subsequent analyses of the research structural model under investigation.

Table 16

Modified Item Linguistic Complexity CFA Standardized and Unstandardized Factor Loadings, Standard Errors, and Confidence Intervals

	Unstandardized Factor Loading	Standardized Factor Loading	SE	$C.I._{95}$
No. Total Words	.81	.72	.10	.52, .92
MLU Words	.37	.62	.13	.37, .87
MLU Morphemes*	1.00	.95	.08	.79, 1.11
Developmental Sentence Score	.82	.72	.07	.58, .86

Note. χ^2 (2) = 1.17, p = .56, *RMSEA* < .001, *CFI* = 1.00.

*Indicates the indicator set as a reference indicator for the scale.

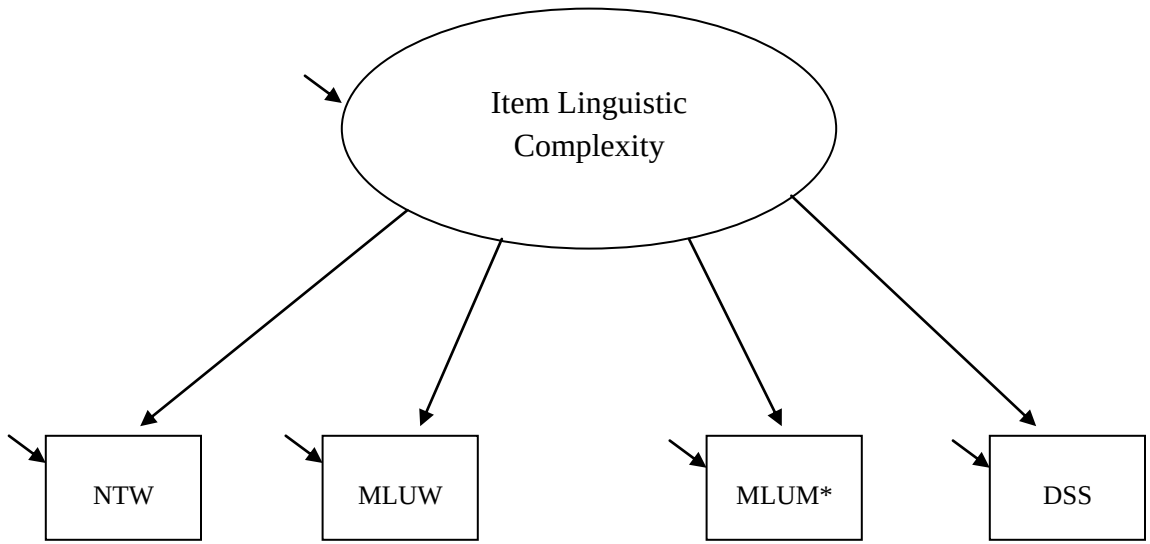

Figure 3. Modified and final item linguistic complexity proposed factor structure.

Note. NTW: number of total words.
MLUW: mean length utterance in words.
MLUM: mean length utterance in morphemes. *As reference indicator.
DSS: developmental sentence score.

Child Language Profile Measurement Invariance

Measurement Invariance Analyses Overview.

Before subsequent testing of the multilevel structural model could proceed, longitudinal measurement invariance was examined using a series of confirmatory factor analyses. Although the *item linguistic complexity* factor remained stable over time (the same standardized items were presented at both assessment time points), the *child language profile* latent factors (at baseline and post intervention) were not assumed to remain invariant over time. A single sample longitudinal approach was used to test the measurement invariance of *child language profile*, in which the baseline measurement structure was compared to the post intervention measurement structure. Robust maximum likelihood estimation (MLV) was performed using Mplus (v.6).

Each time point of *child language profile* was first examined for equal measurement form using a baseline measurement model in which item loadings and intercepts were estimated freely across time points (with the exception of the PPVT scale reference indicators constrained to one). Next, strong measurement invariance was tested by constraining item factor loadings and then the indicator intercepts to equality across time points. However, because the children had received an intervention over the course of the academic year between the baseline and post intervention time points, the criteria for partial measurement invariance were that the majority of factor loadings be invariant across time points. Indicator intercepts were expected to change significantly as a result of intervention.

Equal Forms Measurement Invariance.

The baseline CFA measurement model estimated *child language profile* simultaneously at baseline and post intervention time points with item loadings and intercepts freed across time points, representing a baseline equal forms model in which child language profile was posited to

be structurally stable across time. Because the approach was single sample, longitudinal measurement invariance, indicator and latent factor disturbances were allowed to covary over time (the random measurement error and indicator specific variance was assumed to be temporally stable). *Child language profile* latent factor means were fixed at zero for scaling and identification purposes, such that indicator intercepts were equal to indicator observed means. Figure 4 displays the baseline measurement model.

The chi-square exact fit test indicated significant misfit between the initial baseline model and the data, $\chi^2(95) = 202.49$, $p < .001$. However, approximate fit statistics indicated that the baseline model was an approximate good fit for the data, RMSEA = .07, CFI = .97. All indicators loaded significantly and saliently on the single *child language factor*. At baseline, completely standardized factor loadings ranged from .77 to .84. Post intervention, completely standardized factor loadings ranged from .75 to .86. Indicator variances and correlations were all in the admissible range. The *child language profile* factor at baseline was highly correlated with *child language profile* post intervention, $r = .97$, $p < .001$, indicating that these were indeed the same factors across time. This baseline measurement model was used as the equal forms baseline CFA model for subsequent measurement invariance analyses.

Equal Loadings Measurement Invariance.

Weak measurement invariance was tested by constraining indicator factor loadings to equality across time points (here again, scale reference indicators were constrained to unstandardized factor loadings of one). First, an omnibus equal loadings analysis was conducted. The chi-square exact fit test indicated a significant misfit between the omnibus equal loadings model and the data, $\chi^2(102) = 224.56$, $p < .001$. Approximate fit statistics indicated that this

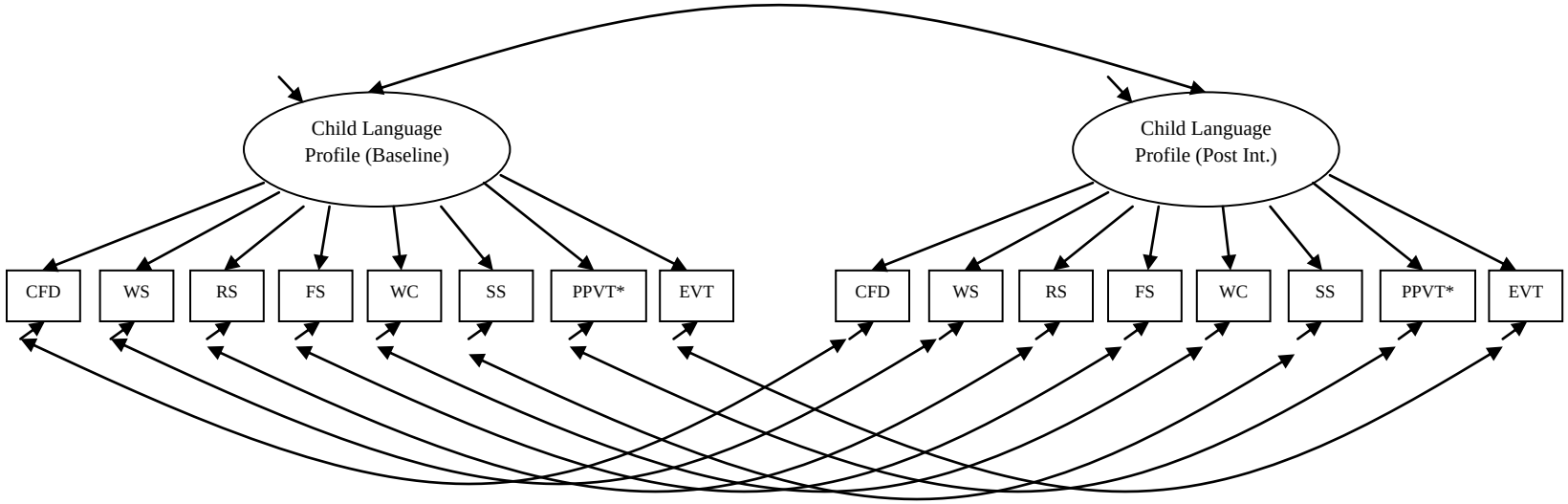

Figure 4. Baseline CFA measurement model for measurement invariance testing. $\chi^2(95)$ = 202.49, p < .001, *RMSEA* = .07, *CFI* = .97.
*Indicates the indicator set as a reference indicator for the scale.
CFD: Concepts and Following Directions subtest of the CELF.
WS: Word Structure subtest of the CELF raw score.
RS: Recalling Sentences subtest of the CELF raw score.
FS: Formulated Sentences subtest of the CELF raw score.
WC: Word Choices I subtest of the CELF raw score.
SS: Sentence Structure subtest of the CELF raw score.
PPVT: Peabody Picture Vocabulary test raw score.
EVT: Expressive Vocabulary Test raw score.

omnibus equal loadings model was an approximate good fit for the data, *RMSEA* = .07, *CFI* = .96; however, chi-square difference testing, comparing this equal loadings model to the more general, baseline equal forms model indicated that the equal loadings model significantly degraded model fit with the data, $\chi^2(6.678) = 22.78$, $p < .001$. Equal loadings testing proceeded to test each indicator loading for temporal equality in a series of individual equal loadings analyses. Table 17 provides global fit statistics and chi-square difference testing results for all models tested.

The indicator factor loadings for the CELF-4 Concepts and Following Directions subtest and the EVT did not appear to be temporally invariant. A final, partial weak measurement invariance model was analyzed, in which all indicator loadings except for the CFD subtest and the EVT were constrained to temporal equality. This partial, weak measurement invariance model was an approximate good fit for the data, $\chi^2(100) = 210.44$, $p < .001$, *RMSEA* = .07, *CFI* = .97, and did not significantly degrade the model fit with the data $\chi^2(4.68) = 7.61$, $p = .11$. Unstandardized and standardized factor loadings, standard errors, and confidence intervals are displayed in Table 18. Given that the majority of factor loadings were invariant across time, and given the approximate good fit of the partial invariance measurement model, this model was considered sufficient for subsequent structural invariance analyses. Figure 5 depicts the final, partial invariance measurement model.

Equal Intercepts Measurement Invariance.

Finally, measurement invariance testing concluded with a test of the intercept equality, by constraining indicator intercepts to equality across time points. Indicator intercepts were not expected to remain stable across time due to the fact that children received intervention and grade-appropriate curriculum instruction over the course of the academic year between baseline

Table 17

Measurement Invariant Child Language Profile Models: Equal Loadings Tests with Global Fit Chi-square Statistics and Chi-square Difference Tests

Model	Chi-Square	DF	Scaling Correction MLR	Adjusted DF for Model	Scaling Correction for Diff test	Satorra-Bentler Scaled Chi-Square Diff Test	p-value
Equal Forms Baseline	202.49	95	1.056	100.32			
Equal Loadings (omnibus)	224.56	102	1.049	106.998	.954	$X^2(6.678) = 22.78$	< .001
Equal Loadings (CFD Constrained Eq)	213.05	96	1.058	101.568	1.248	$X^2(1.248) = 9.28$	< .01
Equal Loadings (WS Constrained Eq)	205.92	96	1.057	101.472	1.152	$X^2(1.152) = 3.32$	.07
Equal Loadings (RS Constrained Eq)	205.24	96	1.058	101.568	1.248	$X^2(1.248) = 2.65$	.10
Equal Loadings (FS Constrained Eq)	203.86	96	1.057	101.472	1.152	$X^2(1.152) = 1.44$	.23
Equal Loadings (WC Constrained Eq)	202.43	96	1.056	101.376	1.056	$X^2(1.056) = .06$	.80
Equal Loadings (SS Constrained Eq)	204.14	96	1.056	101.376	1.056	$X^2(1.056) = 1.65$	.20
Equal Loadings (EVT Constrained Eq)	213.19	96	1.058	101.568	1.248	$X^2(1.248) = 9.39$	< .01
Partial Equal Loadings (CFD & EVT free)	210.44	100	1.050	105	.936	$X^2(4.680) = 7.61$	.11

Table 18

Final, Partial Measurement Invariant Child Language Profile Model Standardized and Unstandardized Factor Loadings, Standard Errors, and Confidence Intervals

	Child Language Profile Baseline					Child Language Profile Post Intervention				
	Mean (SD)	Unstd. Factor Loading	Std. Factor Loading	SE	$C.I._{95}$	Mean (SD)	Unstd. Factor Loading	Std. Factor Loading	SE	$C.I._{95}$
Concepts & Following Directions	12.79 (9.09)	.46	.84	.02	.80, .88	16.78 (10.51)	.53	.86	.02	.82, .90
Word Structure	10.69 (6.64)	.34	.84	.02	.80, .88	14.75 (7.29)	.34	.81	.02	.77, .85
Recalling Sentences	17.61 (14.92)	.70	.78	.03	.72, .84	22.96 (15.38)	.70	.79	.03	.73, .85
Formulating Sentences	11.48 (10.51)	.51	.80	.03	.74, .86	16.18 (11.70)	.51	.74	.03	.68, .80
Word Choices	20.97 (10.86)	.49	.77	.02	.73, .81	26.60 (10.31)	.49	.79	.02	.75, .83
Sentence Structure	14.18 (5.18)	.26	.82	.02	.78, .86	17.05 (5.25)	.26	.84	.02	.80, .88
PPVT*	67.25 (21.96)	1.00	.78	.03	.72, .84	73.43 (20.75)	1.00	.82	.02	.78, .86
EVT	49.48 (10.46)	.50	.80	.03	.74, .86	57.00 (11.96)	.59	.84	.02	.80, .88

Note. $\chi^2(100) = 210.44$, $p < .001$, *RMSEA* = .07, *CFI* = .97.
This model did not significantly degrade the model fit with the data $\chi^2(4.68) = 7.61$, $p = .11$.
*Indicates the indicator set as a reference indicator for the scale.
CELF: Clinical Evaluation of Language Fundamentals 4th edition (Semel, Wiig, & Secord, 2003).
PPVT: Peabody Picture Vocabulary Test III Form A (Dunn & Dunn, 1997).
EVT: Expressive Vocabulary Test (Williams, 1997).

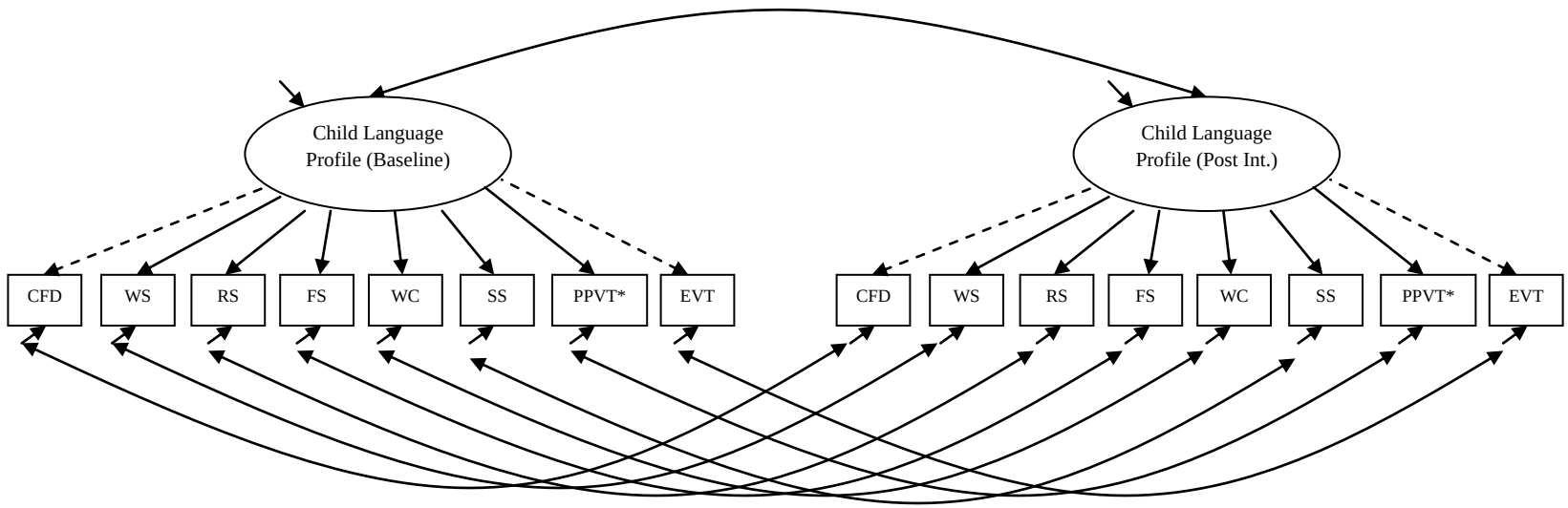

Figure 5. Final, partial invariance measurement model. $\chi^2(100)$ = 210.44, p < .001, *RMSEA* = .07, *CFI* = .97.

Note. *Indicates the indicator set as a reference indicator for the scale.
Dashed lines indicate no equality constraint across time points for these indicators.
CFD: Concepts and Following Directions subtest of the CELF.
WS: Word Structure subtest of the CELF raw score.
RS: Recalling Sentences subtest of the CELF raw score.
FS: Formulated Sentences subtest of the CELF raw score.
WC: Word Choices I subtest of the CELF raw score.
SS: Sentence Structure subtest of the CELF raw score.
PPVT: Peabody Picture Vocabulary test raw score.
EVT: Expressive Vocabulary Test raw score.

and post intervention time points. In fact, children were expected to demonstrate alpha (or true score) changes in mean levels of observed indicators. Intercept equality was tested in an omnibus equal intercepts analysis (with the partial loading invariance established in the previous model). The chi-square exact fit test indicated a significant misfit between the omnibus equal intercepts model and the data, $\chi^2(106) = 508.72$, $p < .001$. Approximate fit statistics also indicated that this omnibus equal intercepts model was a poor fit for the data, *RMSEA* = .12, *CFI* = .88. Chi-square difference testing, comparing this equal intercepts model to the more general, baseline equal forms model indicated that the equal intercepts model significantly degraded model fit with the data, $\chi^2(10.768) = 326.19$, $p < .001$. Measurement invariance testing concluded here, with the final, partially measurement invariant model of *child language profile*. Though the *children's language profile* factors remained structurally stable with largely equal indicator factor loadings over time, the indicator intercepts changed significantly over the course of one academic year.

Multilevel Structural Model Analyses

Structural Model Analyses Overview.

After testing *child language profile* measurement invariance across time points, analyses proceeded to test the hypothesized, multilevel structural model. Multilevel modeling in Mplus (v.6) with Monte Carlo integration was used to analyze the two level interaction between *child language profile* (at the between level) and *item linguistic complexity* (at the within level) in predicting math achievement. Student grade level was treated as a child level covariate. Note that due to the use of a random model type for the within level model (two-level random for subsequent models), only maximum likelihood estimation could be used to estimate models (MLR), and no exact fit statistics, approximate fit statistics, standardized model results, or modification indices were available. Instead, model fit was considered in terms of the number of

free parameters estimated and loglikelihood values. Because the -2 loglikelihood is distributed approximately as a chi-square with degrees of freedom equal to the difference in parameters between models tested, global model fit was considered under these criteria.

A baseline model was used for testing of relevant structural paths, with all structural paths unconstrained across time points using the modified measurement invariant *item linguistic complexity* CFA model at the within (item) level and the partially measurement invariant *child language profile* CFA model at the between (child) level. The baseline structural model is displayed in Figure 6. Within and between level structural paths are considered for the baseline and post intervention models in sections to follow.

Analyses then proceeded to test for structural invariance across time by constraining structural paths of interest to equality one at a time. First, the main effect of *item linguistic complexity* was considered at the within (item) level. Next, the main effect of *child language profile* was considered at the between (child) level. Then, the interaction effect of *item linguistic complexity* and *child language profile* was considered. A final, partially structural-invariant model across time was evaluated against the baseline structural model.

Baseline Structural Model Analysis.

The null hypothesis loglikelihood value (loglikelihood = -64370.86) and MLR scaling correction adjustment were used to generate a chi-square statistic for evaluation of baseline structural model fit. The chi-square statistic indicated significant misfit between the hypothesized structural model and the data, χ^2 *(85.92)* = 128741.72, $p < .001$. Approximate fit statistics were not generated. However, this baseline structural model accepted for additional chi-square difference testing and was judged on structural pathways relevant to the research hypotheses.

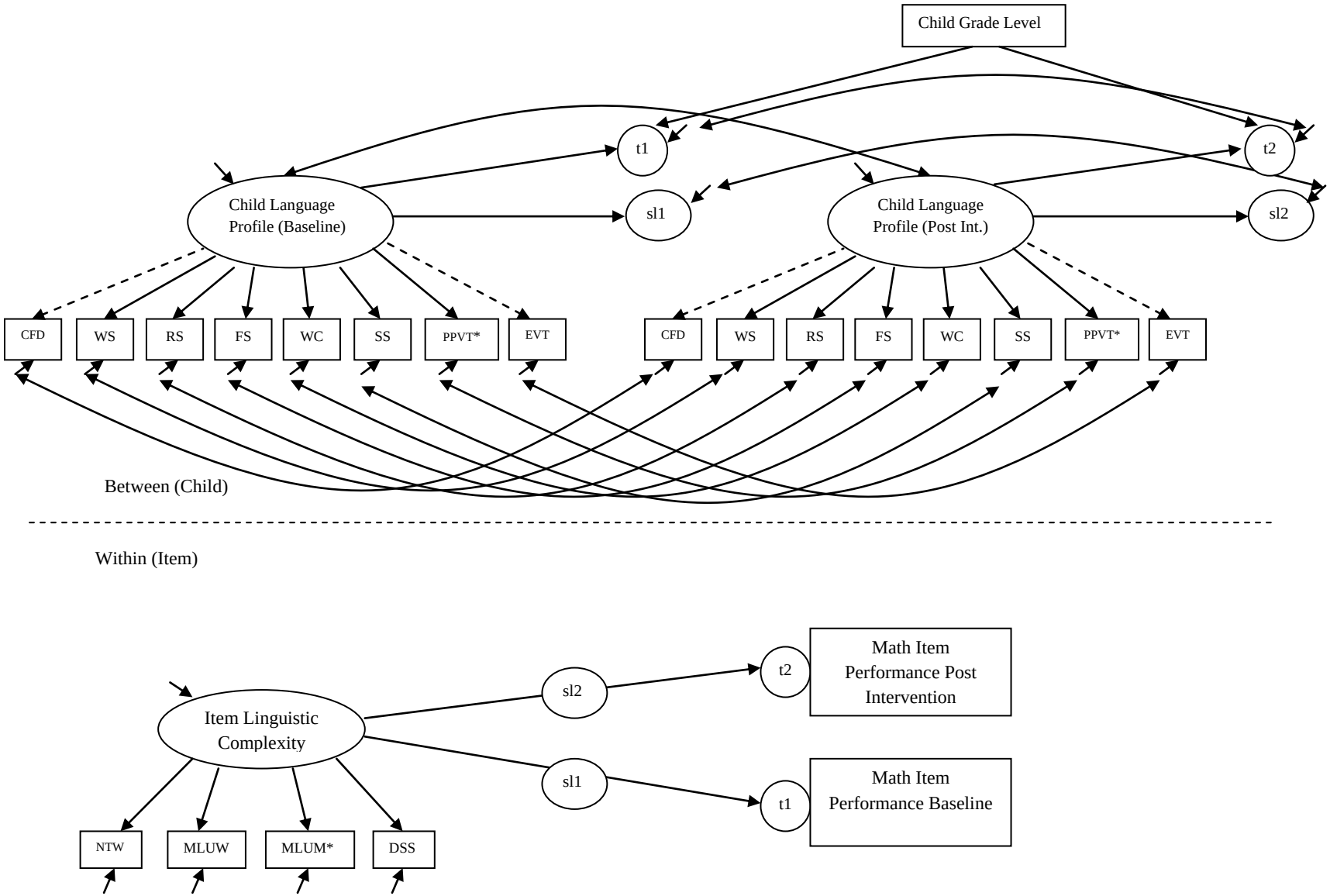

Figure 6. Baseline structural two-level model of mathematics item performance.

The within level portion of the structural model regressed item baseline performance and item post-intervention performance on *item linguistic complexity*, a latent factor indicated by item number of total words, mean length utterance in words, mean length utterance in morphemes, and developmental sentence score. This model created a simple random slope to describe the relationship between *item linguistic complexity* and baseline mathematics item performance (sl1, with random intercept t1) and a simple random slope to describe the relationship between *item linguistic complexity* and post-intervention mathematics item performance (sl2, with random intercept t2). Specific information for the within level measurement model can be found in the *item linguistic complexity* CFA results, displayed in Table 16.

The unstandardized effect of *item linguistic complexity* on baseline item performance was statistically significant and negative, $B = -.07$, $SE = .01$, $p < .001$. This relationship represents the main effect of *item linguistic complexity* on mathematics item performance at baseline, and it can be interpreted to mean that for each one unit increase in latent *item linguistic complexity*, probability of correctly answering a KM-R item decreased by 7% on average. Similarly the unstandardized effect of *item linguistic complexity* on post-intervention item performance was also statistically significant and negative, $B = -.11$, $SE = .01$, $p < .001$. The main effect of *item linguistic complexity* on mathematics item performance post intervention indicated that for every one unit increase in latent *item linguistic complexity*, probability of correctly answering a KM-R item decreased by 11% on average. These main effect structural pathways were compared for statistical equivalence in subsequent structural invariance analyses.

The between level portion of the structural model regressed the random intercepts of *item linguistic complexity* and item performance (*t1* at baseline and *t2* post intervention) on both *child*

language profile (at baseline and post intervention) and child grade level. These pathways represented the main effect of *child language profile* on item performance and the control for grade level as a covariate respectively. Specific information for the between level measurement model can be found in Table 18.

The unstandardized effect of *child language profile* on baseline item performance was statistically significant and positive, $B = .04$, $SE = .004$, $p < .001$. This relationship represents the main effect of *child language profile* on mathematics item performance at baseline, and it can be interpreted to mean that for each one unit increase in latent *child language profile*, probability of correctly answering a Key Math Inventory item increased by 4% on average. Similarly the unstandardized effect of *child language profile* on post-intervention item performance was also statistically significant and positive, $B = .03$, $SE = .003$, $p < .001$. The main effect of *child language profile* on mathematics item performance post intervention indicated that for every one unit increase in latent *child language profile*, probability of correctly answering a KM-R item increased by 3% on average. These main effect structural pathways were compared for statistical equivalence in subsequent structural invariance analyses.

The between level portion of the structural model also regressed the simple random slopes of *item linguistic complexity* and item performance (*sl1* from baseline and *sl2* from post intervention) on *child language profile* latent factors (at baseline and post intervention). These pathways represented the latent interaction between *child language profile* and *item linguistic complexity* in predicting mathematics item performance.

The unstandardized effect of *child language profile* on the relationship between *item linguistic complexity* and baseline item performance was statistically significant and negative, $B = -.001$, $SE < .001$, $p < .001$. This relationship represents the interaction effect of *child language*

profile in predicting the relationship between *item linguistic complexity* and mathematics item performance at baseline, and it can be interpreted to mean that for each one unit increase in latent *child language profile*, the negative effect of *item linguistic complexity* on mathematics item performance was increased, lowering the probability of correctly answering a KM-R item decreased by an additional .10% on average. Similarly the unstandardized effect of *child language profile* on the relationship between *item linguistic complexity* and post intervention item performance was also statistically significant but positive, $B = .001$, $SE < .001$, $p = .046$. The interaction effect of *child language profile* in predicting the relationship between *item linguistic complexity* and mathematics item performance post intervention indicated that for every one unit increase in latent *child language profile*, the negative effect of *item linguistic complexity* on mathematics item performance was lessened, increasing the probability of correctly answering a KM-R item by .10% on average. These interaction effect structural pathways were compared for statistical equivalence in subsequent structural invariance analyses.

Baseline to Post Intervention Structural Invariance Analyses.

First, longitudinal structural invariance of the *item lingusitic complexity* main effect was examined by constraining the baseline and post intervention relationships between *item linguistic complexity* and mathematics item performance to equality. Compared to the baseline measurement model, the constrained *item linguistic complexity* model significantly improved the model fit with the data, *loglikelihood* = -64363.108, χ^2_{diff} *(.84)* = -441.49, $p < .001$ (using the Sartorra-Bentler scaled chi-square difference testing method for MLR). At both baseline and post intervention time points, the main effect of *item linguistic complexity* was equivalent, $B = -.09$, $SE = .01$, $p < .001$. For every one unit increase in latent *item linguistic complexity*, probability of correctly answering a KM-R item decreased by 9% on average.

Next, longitudinal structural invariance of the *child language profile* main effect was examined by constraining the baseline and post intervention relationships between *child language profile* and mathematics item performance random intercepts (*t1* at baseline and *t2* post intervention) to equality. Compared to the baseline measurement model, the constrained *child language profile* model significantly degraded the model fit with the data, *loglikelihood* = -66571.83, χ^2 *(.82)* = 9638.85, $p < .001$. The structural contributions of *child language profile* at baseline and post intervention time points were significantly different. Specifically, although *child language profile* at both time points was a significant and positive predictor of mathematics item performance, the main effect of *child language profile* post intervention was lower in magnitude as a predictor of mathematics achievement.

Finally, the longitudinal structural invariance of the *child language profile* and *item linguistic complexity* interaction effect was examined by constraining the baseline and post intervention relationships between *child language profile* and mathematics item performance random slopes (*sl1* at baseline and *sl2* post intervention) to equality. Compared to the baseline measurement model, the constrained interaction model significantly degraded the model fit with the data, *loglikelihood* = -66214.76.83, χ^2 *(6.59)* = 2550.67, $p < .001$. The structural contributions of the *child language profile* x *item linguistic complexity interactions* at baseline and post intervention time points were significantly different. Specifically, although increases in *child language profile* at baseline increased the negative effect of *item linguistic complexity*, the interaction effect of *child language profile* post intervention was in the opposite direction, decreasing the negative effect of item linguistic complexity on children's probabilities of correctly answering mathematics items. The final, partial structural invariant model is displayed in Figure 7 with path estimates for hypothesized pathways.

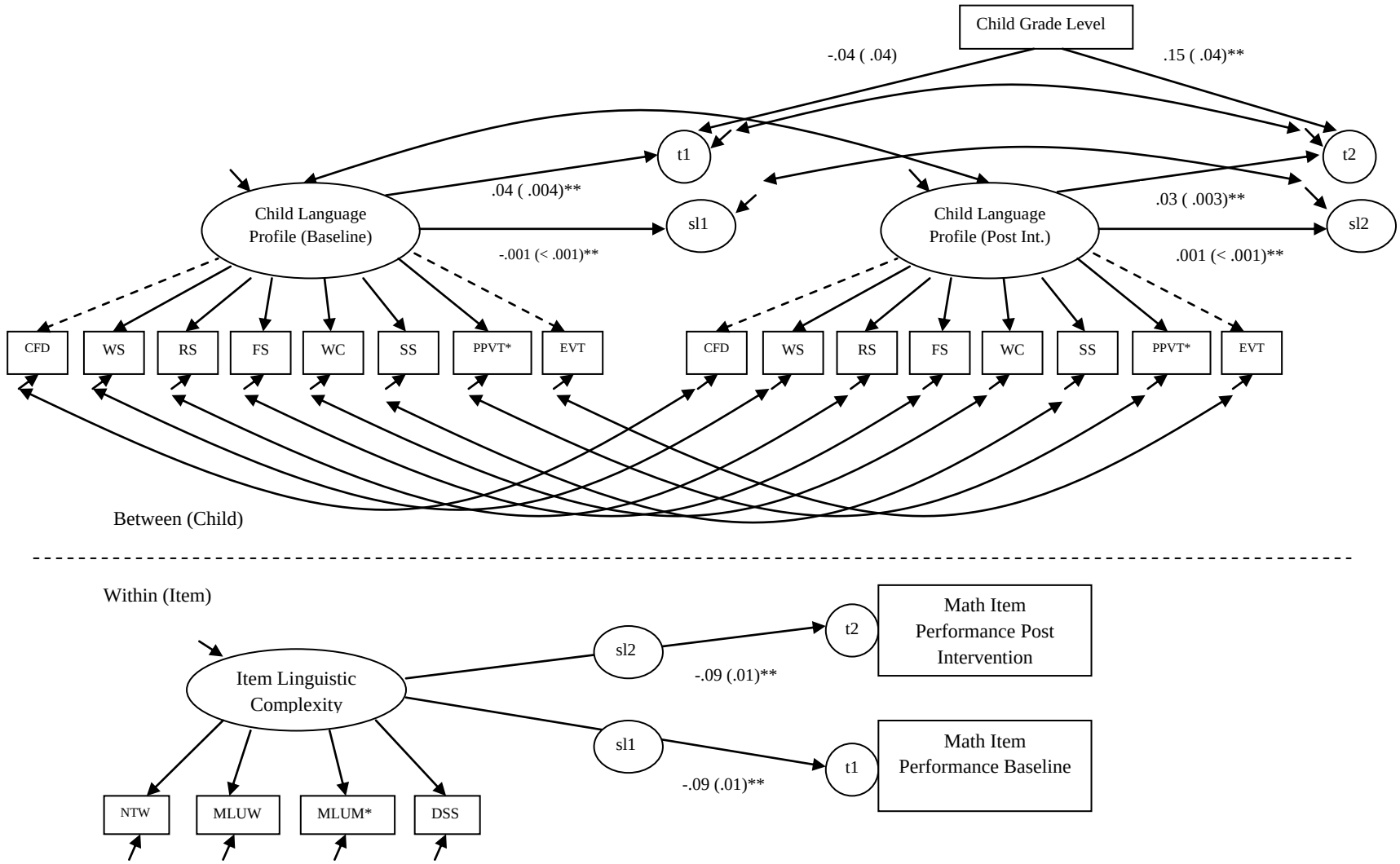

Figure 7. Final, partial structurally invariant, two-level model of mathematics item performance.

Unstandardized effects reported for pathways relevant to research hypotheses. H_0 loglikelihood = -64363.11, free parameters = 79. χ^2 diff (85.08) = -441.49, $p < .001$, significantly improved fit over baseline model.

CHAPTER 4: DISCUSSION

The current study sought to examine the relationship between item linguistic complexity and children's language skills in predicting mathematics achievement of children with mild intellectual disability in terms of (1) a potential main effect of item linguistic complexity, (2) a potential main effect of children's language skills, and (3) a potential interaction between item linguistic complexity and children's language skills. Additionally, this study sought to examine the stability of these predictors of mathematics achievement over time.

At both time points, item linguistic complexity was an important and stable predictor of mathematics achievement. As items increased in linguistic complexity, the probability that children would correctly answer them decreased significantly. Similarly, at both time points, children's language skills were a significant predictor of mathematics achievement. As children's language skills increased, the probability that they would correctly answer mathematics items increased, and the main effect of children's language skills was stronger after intervention than at baseline; however, it is misleading to interpret these main effects in the presence of a significant interaction. The interaction between item linguistic complexity and children's language skills was significant both before and after intervention, but this relationship was not stable across time. Before receiving intervention, children with higher language skills performed worse on linguistically complex mathematics items on average: their language skills increased the negative effect of item linguistic complexity. After receiving intervention, the interaction between these children's language skills and the item linguistic complexity was in the reverse direction. Children with higher language skills performed better on linguistically complex mathematics items: their language skills decreased the negative effect of item linguistic complexity.

The direction of the interaction between children's language skills and item linguistic complexity at baseline was unexpected. If anything, one would expect that language skills would help children interpret the demands of linguistically complex items, regardless of their educational experiences. However, the children's language skills at baseline were quite low, as is often the case for children with mild intellectual disability. The interaction effect at baseline may represent floor effects for children's language skills at this time point or their use of language strategies that were not effective for coping with linguistically complex items.

Interestingly, after receiving intervention with a focus on explicit instruction, children's language skills improved, and the direction of the interaction reversed. Increased language skills helped to curb the stable, negative effect of item linguistic complexity. For each unit increase in children's language skills, the negative effect of item linguistic complexity was reduced by .10% probability that children would provide a correct answer on average. This result would indicate that the relationship between children's language skills and mathematics item linguistic complexity can be altered with targeted intervention. After receiving explicit instruction, children may both improve their language skills and improve their language strategies for approaching linguistically demanding tasks.

Miller, Chapman, and MacKenzi (1981) indicated that language functioning often represents a significant impairment for overall functioning in children with intellectual disabilities. The results of this study support the notion that language is a predictor of functioning in the specific area of mathematics achievement testing for children with mild intellectual disability. As Miller, Chapman, and MacKenzi (1981) indicated, the language impairments of children with MID may mask their general cognitive functioning in other areas, including mathematics achievement.

Previous studies have questioned the content and construct validity of popular mathematics assessments and recommended that tests should be revised to include balanced coverage of mathematics concepts that is relevant to curriculum emphasized at classroom level and in students' IEPs (e.g., Parmar, Fazita, & Cawley, 1996). For children with MID specifically (and most likely for children who have language difficulties in general), the results of the current study indicate that mathematics test revision should also include special considerations for the linguistic complexity of the mathematics assessment items. However, providing testing accommodations which allow for the reading of questions aloud, the repetition of questions prompts, extra time for test completion, and redirections to stay on task (as specified by students' IEPs) may not be enough to help students cope with linguistically complex items. Each of these testing accommodations was allowed in the current study, and the effect of linguistic complexity was still present. Reading aloud, repetition, extra time, and redirection do not change the amount of information that children are asked to store and manipulate to comply with testing demands. Test developers may need to address the linguistic complexity of items during test development, rather than relying on testing accommodations after the fact, if these assessments are to be used for curriculum recommendations with students who have intellectual disabilities.

Limitations and Suggestions for Future Research

The language profile factor used within the current study was broadly defined to include syntax, morphology, vocabulary, and semantics; however, smaller scale features of language may also play a role in predicting the mathematics achievement of children with MID. The phonological loop, a component of verbal working memory which works to temporarily store verbal information, represents a significant deficit for children with MID as compared to typically developing children of the same chronological age and of the same mental age

(Baddeley, 2000; Van der Molen, Van Luit, Jongmans, & Van der Molen, 2007). The amount of verbal information that children with MID can store and process in working memory may be an important predictor of their ability to perform on linguistically complex assessment items. The contribution of other features of language processing such as phonological awareness and verbal working memory should be examined within the language construct in future studies of item linguistic complexity and mathematics performance.

Although this sample of school-aged children with MID was quite large, the number of children in each intervention group (approximately 80) was not large enough for the effects of item linguistic complexity, children's language skills, and their interaction to be examined by group. Overall, it can be said that language-based intervention beyond what was typically offered in special education curriculum was helpful in increasing children's language skills and thereby helping them to cope with linguistically challenging mathematics items. However, the specific components of intervention that impacted this relationship are an area in need of additional research.

The study sample also was drawn from a population of children who had been identified by their schools as having MID and placed in special education programs in the metro-Atlanta area, and generalizing these results to other school systems should be done with caution. School curricula, policies for intelligence testing, and general classroom experiences can vary between individual schools, school systems, and states. Future research should examine the contributions of educational experiences at the classroom, school, and school system levels, with explicit control for IQ (including IQ test results from a valid measure of IQ that have been obtained within two years of a child's inclusion in the study), in an effort to identify environmental variables that may contribute to mathematics achievement.

Although the questions within the KeyMath-Revised Diagnostic Inventory of Essential Mathematics were designed to increase in conceptual difficulty as the test was administered, the linguistic complexity of these questions was not necessarily related to their placement within subtests. Difficulty statistics were provided by the KM-R test developer, but these statistics were based upon a classical testing theory definition of difficulty (proportions of students answering the questions correctly) with a norming sample that did not include students with intellectual disabilities. These difficulty statistics were assumed to be under a unidimensional model of the KM-R, in which mathematics knowledge or ability was the only dimension being tested; however, the results from the current study suggest that for students with MID, linguistic complexity of the mathematics items represents an additional dimension of achievement.

Additional research should examine the relationship between item linguistic complexity and children's language skills with control for an item's mathematical difficulty. This may be best approached from a multidimensional item response theory framework, in which both the mathematics conceptual difficulty and the linguistic complexity of an item are dimensions along which item difficulty can increase. In order for test results to be interpreted for students who are below average in both mathematics and language ability dimensions, a representative sample of children who are not functioning at grade level (including populations of children with MID) should be examined for differential item functioning. Separate norms should be considered.

Similarly, the multidimensional approach to examining the mathematics testing performance of children with MID should move beyond the dichotomous scoring system to include a polytomous scoring system. While a finding of linguistic complexity differentially affecting the likelihood of correctly answering a mathematics item provides support for the idea that children with MID are making errors of understanding, it does not conclusively prove this

point. The dichotomous (right/wrong) scoring system does not allow for a specific characterization of the error(s) contributing to incorrect answers. Additional research should investigate the specific error patterns of this population using approaches such as polytomous scoring, qualitative testing notes, and a dynamic testing methodology.

Conclusions and Practical Applications

Practitioners should use caution in interpreting the mathematics testing performances of children with mild intellectual disability (and this most likely extends to populations of children who experience difficulty with language in general). To some extent, mathematics achievement difficulty may be related to an interaction between the language skills that children bring to testing situations and the linguistic demands of the tests themselves.

If the goal is to make recommendations for classroom placement alone, the traditional approach to standardized mathematics assessments that are dichotomously scored can help practitioners to identify children who are in need of remedial instruction and/or specialized intervention. However, if the goal is to identify the specific areas for intervention or the best approach to intervention, practitioners should be aware that for some children, difficulty with these mathematics items may be related to underlying difficulties with language processing. Specific intervention in mathematics concepts alone may not be enough to remedy achievement difficulties. Language-based intervention also may be needed.

Identification of the specific errors and overall error patterns that are leading to achievement difficulty may be a helpful approach for practitioners seeking to design effective interventions. A dichotomous (right/wrong) scoring system does not allow for a specific characterization of the errors children are making in arriving at the correct answer, but utilizing existing mathematics assessments with a dynamic testing approach and qualitative notes about

performance may be an option for practitioners seeking to evaluate the language confounds of mathematics assessments for children with MID. However, in order for practitioners to truly utilize standardized mathematics achievement assessments to make decisions about intervention designs, test developers may need to evaluate the assumption of unidimensionality (that assessments are only testing math knowledge) for populations of children who have language impairments.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for education and psychological testing*. Washington, DC: American Educational Research Association.

Baddeley, A.D. (2000). "The episodic buffer: a new component of working memory?". *Trends in Cognitive Science* 4: 417–423.

Bhasin, T. K., Brocksen, S., Avchen, R. N., & Van Naarden Braun, K. (2006). Prevalence of four developmental disabilities among children aged eight years -- Metropolitan Atlanta Developmental Disabilities Surveillance Program, 1996 and 2000. *Surveillance Summaries, 55*, 1-9.

Bloom, L. & Lahey, M. (1978). *Language Development and Language Disorders*. New York: John Wiley & Sons.

Boyle, C. A., Yeargin-Allsopp, M., Doernberg, N. S., Holmgreen, P., Murphy, C. C., & Schendel, D. E. (1996). Prevalence of selected developmental disabilities in children 3-10 years of age: The Metropolitan Atlanta Developmental Disabilities Surveillance Program, 1991. *Surveillance Summaries, 45*, 1-14.

Centers for Disease Control and Prevention (1996). State-specific rates of mental retardation -- United States, 1993. *Morbidity and Mortality Weekly Report, 45*, 61-65.

Connolly, A. J. (1988). *KeyMath-Revised: A diagnostic inventory of essential mathematics examiner manual.* Pines, MN: American Guidance Service.

Connolly, A. (1998). *KeyMath-Revised Normative Update: A diagnostic inventory of essential mathematics*. Circle Pines, MN: American Guidance Service.

CTB/McGraw-Hill. (1989). *The comprehensive tests of basic skills*. Monterey, CA: Author.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test – Third Edition*. Circle Pines, MN: American Guidance Service.

Engelmann, S., & Bruner, E. C. (1988a). *Reading Mastery I/II Fast Cycle. Teacher's Guide.* Chicago, IL: SRA.

Engelmann, S., & Bruner, E. C. (1988b). *Reading Mastery I/II. Teacher's Guide.* Chicago, IL: SRA.

Engelmann, S., & Carnine, L. (1992). *Connecting Math Concepts.* Chicago, IL: SRA.

Goodstein, H. A., Kahn, H., & Cawley, J. F. (1976). The achievement of educable mentally retarded children on the Key Math Diagnostic Arithmetic Test. *Journal of Special Education, 10*, 61-70.

Hollingshead, A. B. (1975). *The Hollingshead Two Factor Index of Social Position*. New Haven, CT: Department of Sociology, Yale University.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1993). *The Iowa Tests of Basic Skills*. Chicago: Riverside.

Hu, L. T., & Bentler, P. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural Equation Modeling. Concepts, Issues, and Applications* (pp. 76-99). London: Sage.

Institute of Education Sciences, National Assessment of Educational Progress. (2010). *The nation's report card*. Retrieved from http://nationsreportcard.gov.

Kaufman, A., & Kaufman, N. (1985). *Kaufman Assessment Battery for Children.* Circle Pines, MN: American Guidance Service.

Lee, L. (1974). *Developmental Sentence Analysis*. Evanston, IL: Northwestern University Press.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). *A theory of lexical access in speech production*. Behavioral and Brain Sciences, 22, 1-75.

Lovett, M. W., Lacerenza, L., & Borden, S. L. (2000). Putting struggling readers on the PHAST track: A program to integrate phonological and strategy-based remedial reading instruction and maximize outcomes. *Journal of Learning Disabilities, 33*, 458-476.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk. 3rd Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Miller, J. F., Chapman, R., & MacKenzi, H. (1981). Individual differences in the language acquisition of mentally retarded children. *Proceedings from the second Wisconsin symposium on research in child language.* Madison, WI: University of Wisconsin.

Muthen, L. K., & Muthen, B. O. (2010). *Mplus: Statistical Analysis with Latent Variables (v.6).* Los Angeles, CA: Authors.

National Center for Education Statistics (2011). The nation's report card: Mathematics 2011 (NCES 2012-458). Institute of Education Sciences, U.S. Department of Education, Washington DC.

Owens, R. E., Metz, D. E., & Haas, A. (2007). *Introduction to Communication Disorders: A Lifespan Perspective 3rd edition.* Boston, MA: Pearson Education.

Parmar, R. S. (1992). Protocol analysis of strategies used by students with mild disabilities when solving arithmetic word problems. *Diagnostique, 17*, 227-243.

Parmar, R. S., Frazita, R., & Cawley, J. F. (1996). Mathematics assessment for students with mild disabilities: An exploration of content validity. *Learning Disability Quarterly, 19*, 127-136.

Psychological Corporation (1992). *Weschsler Individual Achievement Test.* San Antonio, TX: Author.

Roeleveld, N., Zielhuis, G. A., Gabreels, F. (1997). The prevalence of mental retardation: A critical review of recent literature (1997) open access. *Developmental Medicine and Child Neurology, 39*, 125-132.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals – Fourth Edition*. San Antonio, TX: Psychological Corporation.

Sevcik, R. A. (2005). Evaluating the effectiveness of reading interventions for students with mild mental retardation. Grant funded by the U. S. Department of Education Institute of Educational Sciences.

STEM Education Coalition. (2000). *Before it's too late: A report to the nation from the national commission on mathematics and science teaching for the 21st century*. Retrieved from http://www2.ed.gov/inits/Math/glenn/report.pdf.

U.S. Department of Education (2009). *2007 annual report to Congress on the implementation of the Individuals with Disabilities Education Act*. Washington, DC: U.S. Government Printing Office.

Van der Molen, M. J., Van Luit, J. E., Jongmans, M. J., & Van der Molen, M. W. (2007). Verbal working memory in children with mild intellectual disabilities. *Journal of Intellectual Disability Research, 51*, 162-169.

Walker, D. W., & Arnault, L. S. (1991). Factorial validity of the KeyMath-Revised. *Diagnostique*, 16, 77-83.

Williams, K. T. (1997). *Expressive Vocabulary Test*. San Antonio, TX: Psychological Corporation.

Williams, T. O. Jr., Fall. A., Eaves, R. C., Darch, C., & Woods-Groves, S. (2007). Factor analysis of the KeyMath-Revised Normative Update Form A. *Assessment for Effective Intervention*, 32, 113 – 120.

Wolf, M., Miller, L., & Donnelly, K. (2000). RAVE-O: A comprehensive fluency-based reading intervention program. *Journal of Learning Disabilities, 33*, 375-386.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri R. (1995). *The Educator's Word Frequency Guide.* Brewster, NY: Touchstone Applied Science Associates.

APPENDICES

Appendix A

The KeyMath-Revised Diagnostic Inventory of Essential Mathematics (Connolly, 1988) features 258 items across 13 subtests and three major concentration areas of mathematics. It is claimed to be diagnostic in part because each of the 13 subtests is theorized to indicate one of the three major mathematical concentration areas. The Basic Concepts area is theoretically indicated by performance in Numeration, Rational Numbers, and Geometry. The Operations area is theoretically indicated by performance in Addition, Subtraction, Multiplication, Division, and Mental Computation. Finally, the Applications area is theoretically indicated by performance in Measurement, Time and Money, Estimation, Interpreting Data, and Problem Solving. However, empirical support (in the form of latent factor analysis) for this three factor model of the KeyMath-Revised is not provided by Connolly (1988) to aid in score interpretation and determination of diagnostic validity.

Walker and Arnault (1991) noted that the factorial validity, and therefore the construct validity, of the KeyMath-Revised Diagnostic Inventory of Essential Mathematics (KM-R; Connolly, 1988) had not been empirically established by the test developer. These authors criticized the Key Math-Revised as a test that had construct validity only established in the areas of developmental skill progression, moderate subtest-total score correlations, moderate subtest-area score correlations, and moderate convergent validity with other popular mathematics instruments (e.g., the Comprehensive Tests of Basic Skills or CTBS and the Iowa Test of Basic Skills or ITBS); however, construct validity in the areas of discriminant validity and factorial validity was not established empirically for the Key Math-Revised. Based on factor analyses of the KM-R total standardization sample intercorrelation matrix, Walker and Arnault concluded

that Connolly's (1988) proposed three factor model for the KM-R was in fact a poor fit for the data, and a two factor model (with allowed dual factor loadings for the Subtraction and Time & Money subtests) was empirically supported. However, these authors noted that the theoretical justifications for the two factor model were not obvious in terms of mathematics skill areas and instead seemed to be a by-product of both item content overlap and formatting issues. Walker and Arnault cautioned diagnosticians against (1) assuming construct validity for the KM-R, and (2) using Connolly's (1988) proposed KM-R factor structure to interpret examinee scores. These authors pointed to the test's name and its popularity as major contributing factors for its wide use in testing school-age children despite a lack of factorial validity.

Despite the critiques and recommendations of Walker and Arnault (1991) the KM-R has remained a popular measure of mathematics, used in diagnosis of mathematics difficulty, educational testing and placement, program planning, and mathematics research. No other research was published on the KM-R factorial validity (and its implications for construct validity) until Williams, Fall, Eaves, Darch, and Woods-Groves (2007) attempted to replicate Walker and Arnault's (1991) KM-R findings with the KeyMath-Revised Diagnostic Inventory of Essential Mathematics Normative Update (KM-R-NU; Connolly, 1998), an updated version of the KM-R with the same 258 items and the same 13 subtests as the 1988 KM-R. Williams et. al. (2007) cited Walker and Arnault's (1991) critique of the earlier version of KeyMath, the continued lack of factorial validity in KM-R considerations of construct validity, and adherence to the measurement standards recommended by the *Standards for Educational and Psychological Testing* (see Standard 1.11; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) as major concerns about the KM-R-NU.

Williams et. al. replicated Walker and Arnault's (1991) findings with a unique sample of 130 children from both public and private schools in the Southeastern United States, who were majority White, balanced for gender, and ranging in grade level from 1st to 12th (M = 6.31, SD = 2.33). This sample included 14 children who were receiving special education services as a result of identification with one or more disabling conditions, including emotional behavioral disorder, learning disability, unspecified disability, and intellectual disability. Williams et al. (2007) used confirmatory factor analysis to examine Connolly's (1988; 1998) three factor model of the KM-R, and found a significant misfit between the three factor model and the data, mediocre to acceptable approximate fit statistics, and collinearity (all factor *r*s > .90) between the three latent factors proposed by Connolly (1998). These authors tested additional models for the KM-R factor structure using exploratory factor analysis, and based on patterns of item loadings and theoretical underpinnings of the KM-R-NU, concluded that a single factor solution, indicating overall mathematics skill, was most appropriate for the KM-R-NU. Williams et. al. (2007) recommended that practitioners avoid using KM-R-NU area scores proposed by Connolly (1988; 1998) and instead base interpretations of KM-R scores on total score performance, as total scores tend to be more robust and were empirically supported by their results.

Despite the statistical quality and practical significance of these factor studies, both the Walker and Arnault (1991) and the Williams et. al. (2007) articles had received relatively little attention from the psychometric and educational achievement communities by the time of this research study, with citations by only three and only one other articles respectively.

The current research study used the KeyMath-Revised Diagnostic Inventory (1988) Form A as a measure of mathematics achievement; however, due to the limited instructional experiences of the sample of children identified as having mild intellectual disability and

receiving special education services, certain subtests of the KeyMath-Revised were not used. Furthermore, due to concerns about limited educational experiences, lack of representation in standardization samples, and interpretability of standard scores, raw scores were used to characterize subtest level KeyMath-Revised performance for the purposes of correlation and covariate analyses. The dependent variable of interest for the proposed study was item level KeyMath-Revised performance. It was anticipated that the concerns about factorial validity of the KeyMath-Revised assessment reflected in research to date were not of direct consequence to these analyses, and that the investigation of item linguistic complexity in predicting mathematics performance might add to the body of literature characterizing the construct validity of this popular mathematics assessment.

Appendix B

Intervention Group Composition Analyses.

One-way ANOVAs and *post hoc* Bonferoni tests were performed to examine group differences across age, grade level, PPVT language age, IQ, and socioeconomic status variables (measured at baseline before intervention). The assumption of homogeneity of variance was met for all aforementioned variables with the exception of grade level (Levene statistic (2, 241) = 15.31, p < .001). Although ANOVA is generally robust to violations of homogeneity of variance with equal sample sizes, because of slight differences in sample size between groups (n = 77 for Math, n = 80 for RAVE-O, and n = 87 for PHAB), the Brown Forsythe nonparametric test was employed to assess group differences for grade level.

The Math, RAVE-O, and PHAB intervention groups were comparable in IQ, $F(2,203) = .768$, $p = .47$, mother occupation score, $F(2,211.91) = .471$, $p = .63$, father occupation score, $F(2,144) = 1.73$, $p = .18$, mother Hollingshead score, $F(2,213) = .85$, $p = .43$, father Hollingshead score, $F(2,139) = 2.38$, $p = .10$, and overall family Hollingshead score, $F(2,221) = 1.67$, $p = .19$. However, the groups were significantly different with respect to age, $F(2,241) = 14.29$, $p < .001$, grade level, $F(2,221.82) = 20.53$, $p < .001$, PPVT language age, $F(2,241) = 5.70$, $p < .01$, and mother education, $F(2,223) = 7.29$, $p = .001$, and the groups were marginally significantly different with respect to father education, $F(2,152) = 3.06$, $p = .05$.

Bonferoni post hoc analyses revealed that although the PHAB and RAVE-O intervention groups were not statistically different in age, the Math intervention group was significantly younger than both (*mean difference* = -8.64, *SE* = 2.40, $p = .001$ and *mean difference* = -12.87, *SE* = 2.45, $p < .001$ respectively). Similarly, although the PHAB and RAVE-O group were not statistically different in grade level or PPVT language age, the Math group was significantly

lower than both PHAB and RAVE-O groups in grade and PPVT language age (grade level *mean difference* = -.85, *SE* = .17, $p < .001$, and *mean difference* = -.99, *SE* = .17, $p < .001$ respectively; PPVT language age *mean difference* = -.69, *SE* = .25, $p = .02$, and *mean difference* = -.79, *SE* = .26, $p = .007$ respectively). The Math intervention group also displayed lower mother education scores than both the PHAB and RAVE-O groups (*mean difference* = -1.59, *SE* = .48, $p = .003$, and *mean difference* = -1.61, *SE* = .49, $p = .003$ respectively; again, PHAB and RAVE-O were not statistically significantly different in mother education levels). In terms of father education levels, although the Math group was not statistically different from the PHAB group, the level of father education in the Math group was significantly lower than the RAVE-O group (*mean difference* = -1.78, *SE* = .72, $p = .04$; again, the PHAB and RAVE-O groups were not significantly different).

Due to these group differences in age, grade level, mother education, and father education, these variables of special interest in subsequent covariate analyses. PPVT language age is predicted by PPVT scores (an IV of interest in the child language profile characterization), and so, PPVT language age is described here in the interest of sample characterization. It was not treated as a covariate in subsequent analyses due to issues of collinearity with the PPVT score as an indicator of *child language profile* latent factor.

Crosstabs and Chi square statistics were performed to examine group compositions across sex, race/ethnicity, school, and county. The groups were comparably distributed in terms of sex, $\chi^2(2) = 2.39$, $p = .30$, race/ethnicity, $\chi^2(8) = 10.92$, $p = .21$, and county, $\chi^2(2) = 1.15$, $p = .56$. The intervention groups were significantly different in their distribution across schools, $\chi^2(22) = 83.21$, $p < .001$. Ideally, in this situation, school could be treated as a level within the planned multilevel analysis. However, with only 12 participating schools, a multilevel analysis

with an additional level for schools would be severely underpowered. The comparability of schools in terms of Title I status, operating budgets, and student demographics are considered in Appendix C.

Appendix C

School characterization across intervention groups.

While school demographic information was readily available for all five years of the parent study, school budget and expenditure reports were not available for fiscal years 2005 - 2006 and 2006 - 2007 (corresponding to years one and two of the parent study). Georgia school expenditure reports include information about numbers of full-time enrolled students, school operating budgets, and school costs of attendance per student. In analyzing for group similarities across school level variables of interest, school fiscal year reports were matched to student years of intervention participation. However, because no spending reports were available for years one and two of the study, school level data was counted as missing for these participants. (Under other circumstances imputation methods such as mean imputation or multiple imputation may be appropriate; however, with only three years of school financial information available and large fluctuations in school financial profiles between years, it was decided that imputation was inappropriate.) Due to the extent of school missing financial data for years one and two of the study, school economic climates were considered using the two variables reported for all 12 schools for all five years of the parent study: Title I status and percentages of students eligible for free or reduced lunch.

Crosstabs and Chi square statistics were performed to examine group compositions across Title I status school membership. A significant association between intervention group assignment and Title I school status was found, $\chi^2(2) = 26.89$, $p < .001$, *Pearson contingency coefficient* = .32. Specifically, for the math intervention group, observed frequencies of membership in Title I schools was greater than expected; for the PHAB intervention group, observed frequency of membership in Title I schools was comparable to expected frequency; and

for the RAVE-O intervention group, observed frequency of membership in Title I schools was less than expected. The Title I membership contingency table is displayed in Table 19.

One-way ANOVAs and *post hoc* Bonferoni tests were performed to examine intervention group differences across percent of students eligible for free or reduced lunch, percent of students with disabilities, percent of Black students, percent of White students, percent of Hispanic students, percent of Asian students, percent Multiracial students, and percent of Native American students. The assumption of homogeneity of variance was met for all aforementioned variables with the exception of percent of Hispanic students, Levene statistic (2, 241) = 18.47, $p < .001$, percent of Asian students, Levene statistic (2, 241) = 11.94, $p < .001$, and percent of Native American students, Levene statistic (2, 241) = 4.14, $p = .02$. Again, although ANOVA is generally robust to violations of homogeneity of variance with equal sample sizes, because of slight differences in sample size between groups (n = 77 for Math, n = 80 for RAVE-O, and n = 87 for PHAB), the Brown Forsythe nonparametric test was employed to assess group differences for variables not meeting the homogeneity of variance assumption.

With consideration for school level variables, the Math, RAVE-O, and PHAB intervention groups were comparable in school percentages of Black students, $F(2,241) = .05$, $p = .95$, school percentages of Asian students $F(2, 225.88) = 1.79$, $p = .17$, school percentages of Native American students $F(2, 232.49) = 1.01$, $p = .37$, and school percentages of students with disabilities, $F(2,241) = 2.02$, $p = .14$. However, the intervention groups were significantly different with respect to school percentages of students eligible for free or reduced lunch, $F(2, 241) = 6.24$, $p = .002$, school percentages of White students, $F(2, 241) = 4.39$, $p = .01$, school percentages of Hispanic students, $F(2, 206.37) = 5.20$, $p = .006$, and school percentages of Multiracial students, $F(2,241) = 5.58$, $p = .004$.

Table 19

Title I Contingency Table

		Title I Status		
		Title I School Participants Observed frequency (Expected frequency)	Non-Title I School Participants Observed frequency (Expected frequency)	Totals
Intervention Group	Math	58 (43.9)	19 (33.1)	77
	PHAB	53 (49.6)	34 (37.4)	87
	RAVE-O	28 (45.6)	52 (34.4)	80
	Totals	139	105	244

Note. χ^2 (2) = 26.89, $p < .001$, *Pearson contingency coefficient* = .32.

In terms of racial/ethnic composition, Bonferoni post hoc analyses revealed that although the PHAB and Math intervention groups were not statistically different from each other in school percentages of White students, on average, the RAVE-O group had significantly higher percentages of White students than both (PHAB *mean difference* = 7.24%, *SE* = 2.69, *p* = .02, and Math *mean difference* = 6.76%, *SE* = 2.78, *p* = .047). Similarly, although the PHAB and Math groups were not statistically different in school percentages of Hispanic students, the RAVE-O group had significantly lower school percentages of Hispanic students than both PHAB and Math groups on average (PHAB *mean difference* = -7.77%, *SE* = 2.64, *p* = .01, and Math *mean difference* = -7.07%, *SE* = 2.72, *p* = .03). While the Math intervention group was not significantly different from either PHAB or RAVE-O groups in school percentages of Multiracial students, the RAVE-O group had significantly lower school percentages of Multiracial students than the PHAB group on average (*mean difference* = -.72%, *SE* = .22, *p* = .003. In summary, the racial/ethnic climates of the schools do seem to vary across intervention groups; however, because the racial/ethnic composition of intervention groups was not significantly different, race/ethnicity was not treated as a covariate in subsequent analyses.

Bonferoni post hoc analyses revealed that although PHAB intervention group was not statistically different from the Math intervention group in school percentages of students eligible for free or reduced lunch, on average, the RAVE-O group had significantly lower school percentages of students eligible for free or reduced lunch (PHAB *mean difference* = -8.64%, *SE* = 3.21, *p* = .02, and Math *mean difference* = -11.02%, *SE* = 3.31, *p* = .003). This finding, combined with the Title I status differences across intervention groups, would seem to suggest that school level economic climate should be controlled for in subsequent multilevel analyses.