AMERICAN JOURNAL ON INTELLECTUAL AND DEVELOPMENTAL DISABILITIES
2015, Vol. 120, No. 6, 542–568

DOI: 10.1352/1944-7558-120.6.542

Testing Math or Testing Language? The Construct Validity of the KeyMath-Revised for Children With Intellectual Disability and Language Difficulties

Katherine T. Rhodes, Lee Branum-Martin, Robin D. Morris, MaryAnn Romski, and Rose A. Sevcik

Abstract

Although it is often assumed that mathematics ability alone predicts mathematics test performance, linguistic demands may also predict achievement. This study examined the role of language in mathematics assessment performance for children with intellectual disability (ID) at less severe levels, on the KeyMath-Revised Inventory (KM-R) with a sample of 264 children, in grades 2–5. Using confirmatory factor analysis, the hypothesis that the KM-R would demonstrate discriminant validity with measures of language abilities in a two-factor model was compared to two plausible alternative models. Results indicated that KM-R did not have discriminant validity with measures of children's language abilities and was a multidimensional test of both mathematics and language abilities for this population of test users. Implications are considered for test development, interpretation, and intervention.

Key Words: *intellectual disability (ID); language; mathematics assessment performance; KeyMath-Revised; discriminant validity; multidimensional assessment*

General mathematics skills are an important aspect of successful daily living (STEM Education Coalition, 2000). 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 students who are not performing at grade level. Although it often has been assumed that poor mathematics test results indicate poor development of mathematics concepts, linguistic demands have rarely been evaluated as potentially confounding assessment effects (Abedi, Hofstetter, Baker, & Lord, 2001; Abedi & Lord, 2001; Abedi, Lord, & Hofstetter, 1998; Abedi, Lord, & Plummer, 1997; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006). The linguistic demands of mathematics assessment may be particularly important for populations of children who experience difficulty with language acquisition and processing (e.g., children with intellectual disability [ID] at less severe levels). To the extent that popular mathematics assessments are linguistically demanding, they may become assessments of the language skills of children with ID and language difficulties; and this issue of test validity is an area in need of additional research.

Language and Mathematics Abilities

Although there is no universally accepted theory of cognitive abilities, Cattell-Horn-Carroll (CHC) theory is perhaps the most widely accepted, unifying theory of cognition to date (Schneider & McGrew, 2012). CHC theory merges Horn and Cattell's *Gf-Gc* theory of fluid and crystallized intelligence (see, e.g., Horn & Cattell, 1966) with Carroll's (1993) three stratum theory of general intelligence, broad abilities, and narrow abilities; and CHC thereby represents an integration of the last century of theory and empirical testing on human cognitive abilities (Schneider & McGrew, 2012).

CHC theory is hierarchically organized such that a general intelligence factor, *g*, predicts

DOI: 10.1352/1944-7558-120.6.542

domain-free general capacities (like fluid intelligence and short-term memory), as well as domain specific acquired knowledge and sensory-motor abilities (like verbal knowledge, quantitative knowledge, and auditory processing), which in turn predict narrow abilities (like language development and mathematical achievement). Overlap (correlation) between factors at the broad ability level is allowed, as is overlap between factors at the narrow ability level; however, they remain distinct constructs which should evidence divergent validity. An abbreviated factor model of CHC theory is presented in Figure 1.

The intelligence factor (*g*) and its explanations remain somewhat controversial, and despite the consensus of CHC theorists regarding broad and narrow cognitive abilities, each theorist had a slightly different interpretation of *g* (see Schneider & McGrew, 2012 for a summary). From the simplest perspective, *g* represents little more than a positive manifold, or a tendency for tests of cognitive ability to correlate moderately and positively. From a more complex perspective, *g* represents an intelligence factor with explanatory power, predicting ability and individual differences across cognitive domains.

The relationship between the CHC theory construct *g* and the measurement of IQ is something that must be considered with respect to a particular intelligence test and the cognitive theory used to guide its design; different tests employ different theories of intelligence and construct definitions for *g*. However, CHC theory is quite ubiquitous in the conceptualization and design of intelligence tests (e.g., the Kaufman Assessment Battery for Children, The Woodcock Johnson tests of Cognitive Abilities, the Stanford-Binet Intelligence Scales; the Wechsler Intelligence Scale for Children, and many more; Naglieri & Goldstein, 2009).

Despite the fact that many popular intelligence tests use CHC theory as a basis for their conceptualization, tests differ in their inclusion of various subscales, representing various broad and narrow ability domains of *g*. These subscales (and the broad and narrow domains they represent) are generally used to form composite, full scale IQ scores, representing *g*. However, regardless of the assessment used and the subscales incorporated in measuring *g*, measures of intelligence generally correlate well with each other if they are measuring *g* (Naglieri & Goldstein, 2009). Thus, we can be reasonably sure that the IQ (measurement of *g*) across various intelligence tests is reflecting the same construct even though different tests focus on different aspects (broad and narrow abilities) composing *g*.

For children with ID, an IQ range of 55 to 70, in conjunction with impairments in adaptive functioning, is often used to determine a diagnosis of ID at less severe levels (Naglieri & Goldstein, 2009). In general, CHC theory would characterize children with ID as having a lower than average general intelligence which affects their development and functioning across a variety of domains; however, children with ID have not been characterized by a traditional pattern of specific subtest (broad or narrow ability) performance (Naglieri & Goldstein, 2009). To the contrary, in addition to exhibiting a general pattern of lower than average performance across a variety of broad abilities, children with ID may also be characterized by heterogeneous performances (sometimes termed "splinter skills") across broad abilities and a variety of ability profiles (Bergeron & Floyd, 2006). Their low global performance (*g*) does not necessarily imply a consistent, low performance in all broad abilities, and as a population, they are characterized by a variety of strengths and weaknesses in broad abilities (Bergeron & Floyd, 2006).

The assumption that all children with ID at less severe levels should invariably display deficits across all broad abilities, including the broad ability domain of *Quantitative Knowledge* (*Gq*) is unqualified, and determining the measurement validity of mathematics achievement instruments that are often used with this population is an area in need of research. Further, regardless of one's particular theoretical alignment regarding *g*, and regardless of any particular individual's general intelligence, narrow abilities like language development and mathematical achievement should remain distinct constructs; and valid tests of these abilities should evidence divergent validity.

Language is commonly understood as a combination of skills in the areas of syntax, morphology, vocabulary (including expressive and receptive vocabulary knowledge), semantics, and pragmatics (Bloom & Lahey, 1978). Indeed, under the widely accepted framework of CHC theory, language development is defined as the core of the unitary factor *Comprehension-Knowledge* (*Gc*), indicated by general verbal information, language development, lexical knowledge, listening ability, communication ability, and grammatical sensitivity (Schneider & McGrew, 2012).

©AAIDD
DOI: 10.1352/1944-7558-120.6.542

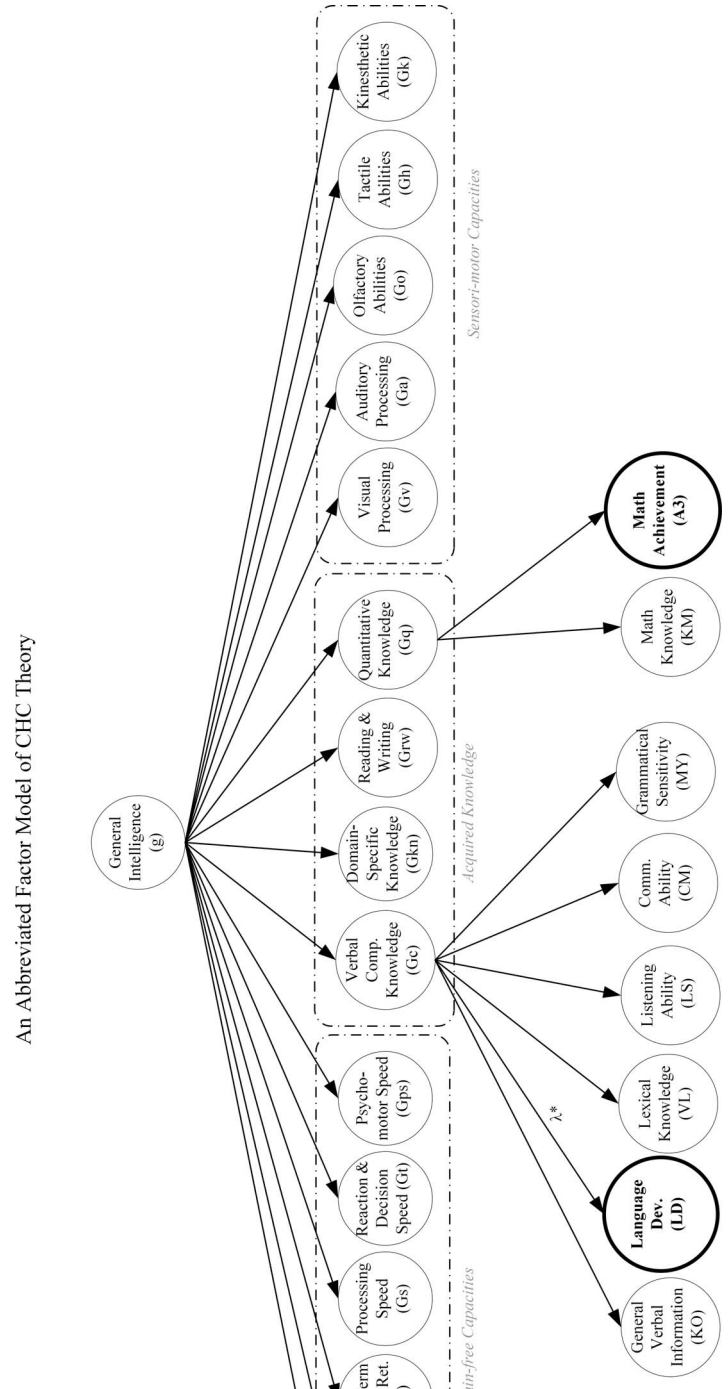

Figure 1. An abbreviated, schematic factor diagram of Cattell-Horn-Carroll (CHC) theory for language and mathematics abilities. For further details on additional abilities, see Schneider & McGrew (2012).

DOI: 10.1352/1944-7558-120.6.542

Empirically, language appears to be a unitary construct when measured by oral and listening vocabulary assessments and tests of listening comprehension; however, the components of language may have differing contributions depending on ages of development (Carroll, 1993). 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; however, these features are perhaps best understood as components of higher order factors such as general intelligence or crystallized intelligence (acquired knowledge), which are predictors of language (Carroll, 1993).

Under CHC theory, *mathematics* is also defined as a unitary construct in the broad cognitive domain of *Gq*, indicated by the narrow abilities of mathematical knowledge and mathematical achievement (Schneider & McGrew, 2012). Currently, CHC theory allows quite a bit of latitude for the validation and addition other narrow abilities to the *Gq* domain (e.g., number sense), and narrow abilities from other broad cognitive domains may also contribute to mathematical cognition (e.g., quantitative reasoning from the broad domain of *Fluid Reasoning*, *Gf*, and numerical facility from the broad domain of *Processing Speed*, *Gs*; Schneider & McGrew, 2012). The *Gq* domain is typically measured with tests of mathematical academic achievement, which are hopefully reflective of students' curricular experiences. For the purposes of the current study, the construct of *Gq* will be referred to as *mathematics*.

Both *language* and *mathematics* fall under CHC theory's broad cognitive domains of acquired knowledge (the additional domains of acquired knowledge are *reading and writing* and *domain-specific knowledge* or specialized knowledge). *Language, reading and writing,* and *specialized knowledge* are theorized to be distinct from the domain of *mathematics*. Therefore, measures of each construct should demonstrate discriminant validity between these three broad domains.

However, *mathematics* is often confounded with abilities of *reading* and *writing* because of test formatting issues. In particular, mathematics tests that are delivered in written formats tap into students' reading abilities, and mathematics tests that require students to return their answers in writing tap into students' writing abilities (Carroll, 1993). Some *mathematics* assessments have moved away from reading and writing formats in order to address questions of confounding, frequently using oral language for the delivery of *mathematics* word problems. Although moving away from written formats may not create assessment difficulty for students with age-appropriate levels of *language* ability, relying on oral *language* to deliver *mathematics* assessment items may be problematic for students with disabilities because of difficulty understanding linguistically complex items (Shaftel et al., 2006).

The extent to which *language* abilities and language-formatted *mathematics* measures are related to each other for students with disabilities is an area in need of additional research. From the most conservative perspective, *language* and language-formatted *mathematics* measures may demonstrate the expected pattern of discriminant validity, yielding separate factors defined by their separate prediction of *language* and *mathematics* items respectively. From the other extreme, *language* may be the sole predictor of language-formatted mathematics item performance for populations with impairments in *language* ability. For some special populations, then, it is possible that *language* is the largest determinant of linguistically delivered items, with no discriminant validity for measuring other abilities such as mathematics.

The current study examined the role of cognitive linguistic skills in the mathematics performance of children with ID at less severe levels, on a popular, language-formatted mathematics skill assessment. The purpose of this research was to characterize the extent to which language skills predict mathematics performances on a selection of items from the KeyMath-Revised Diagnostic Inventory of Essential Mathematics (KM-R; Connolly, 1988) for children with ID through examination of the KM-R factor structure.

The KeyMath-Revised Inventory

Although the KM-R is designed for use with children from a variety of backgrounds and cognitive profiles, it is one of the most widely used mathematics assessments for children with disabilities receiving special education services (Parmar, Frazita, & Cawley, 1996; Walker & Arnault, 1991). The KM-R features 258 items across 13 subtests and three major concentration areas of mathematics. It is thought to be

DOI: 10.1352/1944-7558-120.6.542

diagnostic in part because each of the 13 subtests is theorized to indicate one of the three major mathematical concentration areas: Basic Concepts, Operations, and Applications. The KM-R is widely used, partially because it provides test administrators with norm-referenced performance reports in content-specific areas. Despite these strengths, many researchers have raised concerns about its construct validity, content validity, and formatting. These issues are considered in the sections that follow.

Construct Validity of the KeyMath-Revised

Factor validity of the KM-R: Three versus one. There are three major studies guiding the use and interpretation of the KM-R. These studies include the original development of the test (Connolly, 1998) and two-factor validity studies (Walker & Arnault, 1991; Williams, T. O., Fall, Eaves, Darch, & Woods-Groves, 2007). Although the original study argued for a three-factor structure for the KM-R (Basic Concepts, Operations, and Applications), the additional studies have questioned the construct validity of interpreting the test as representing three factors.

During test development, the 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). The Basic Concepts, Operations, and Applications concentration area correlations ranged from moderate to extremely high (.68 to .92) depending on the ages of examinees tested. However, the factor validity and discriminant validity of the KM-R were not reported or empirically examined during test development. Connolly (1988) ultimately proposed a three-factor structure for the KM-R.

The first psychometric reexamination of the KM-R used the total standardization sample intercorrelation matrix to conclude that the proposed three-factor model for the KM-R was in fact a poor fit for the data (Walker & Arnault, 1991). Instead, Walker and Arnault (1991) found that 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 (a) assuming construct validity for the KM-R, and (b) using Connolly's (1988) proposed KM-R factor structure to interpret examinee scores.

A second study 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), an updated version of the KM-R with the same 258 items and the same 13 subtests as the 1988 KM-R. T. O, Williams et al. (2007) 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 ($M = 6.31$, $SD = 2.33$). These authors found mediocre to acceptable fit of a three factor model to the data and substantial overlap among the three factors (all rs $> .90$). These authors tested additional models for the KM-R factor structure using exploratory factor analysis, and concluded that a single factor solution, indicating overall mathematics skill, was most appropriate for the KM-R. T. O. 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 single factor results.

Discriminant validity: The role of language. Additionally, discriminant validity, which can be defined as the extent to which a test is not highly correlated with tests designed to measure theoretically different constructs (Allen & Yen, 2002), was not examined by KM-R developers (Walker & Arnault, 1991). Examinations of the discriminant validity of the KM-R might have included comparisons with any number of assessments that do not purport to indicate mathematics skill development (e.g., reading or oral language tests); however, given that the primary modality of test question delivery is language-based, establishing discriminant validity with language assessments can be seen as crucial to understanding the construct validity of the KM-R.

Both the factor validity and the discriminant validity of the KM-R are important to empirically examine the construct of *mathematics* skill oper-

DOI: 10.1352/1944-7558-120.6.542

ationalized with this popular mathematics achievement measure. The factor structure of the KM-R will be informative for the discriminant validity of the KM-R with measures of language skill. A failure of the KM-R to demonstrate discriminant validity with measures of language skill could indicate that this test may not measure mathematics performance in a clear and consistent manner. The extent to which the KM-R is a unique indicator of *mathematics* ability, as opposed to an indicator of *language* ability, is a crucial question for the current research.

Content Validity of the KeyMath-Revised

Beyond the issues of KM-R construct validity, the content validity of the KM-R has been questioned for populations of children with ID (Parmar et al., 1996). Content validity for the KM-R was originally 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 (Connolly, 1988). However, this assessment was developed and normed on a sample of 1,794 typically developing students between 5 and 15 years of age. The content validity of the KM-R, when used with populations of children with ID at less severe levels, has been called into question for (a) failure to provide balanced coverage of mathematics concepts appropriate for this population, (b) overemphasis on computation and underemphasis on strategy and problem solving, and (c) mismatch with students' special education classroom experiences and IEP goals (Parmar et al., 1996). These authors noted that for the KM-R (and a number of other popular mathematics achievement assessments), testing recommendations may have little practical relevance to educational placement, curriculum design, and instructional strategies of children with ID. The mathematics content of the KM-R may be appropriate for typically developing children in mainstream learning environments, but it may be problematic for children with ID and are in special education environments.

KM-R Formatting: The Role of Language in Predicting KM-R Performance for Children With ID at Less Severe Levels

Students with ID have been largely excluded from the developmental research on mathematics difficulty to date. Most of the developmental research on the mathematics performance of students with disabilities is focused on students with learning difficulties or learning disabilities (see, e.g., Geary, 1993; Lyon, Shaywitz, & Shaywitz, 2003; Mazzocco & Myers, 2003; Swanson & Jerman, 2006)–not necessarily students with ID. In general, large scale national achievement testing indicates that the majority of students with disabilities, including students with ID, do not reach grade level proficiency in mathematics (U.S. Department of Education, 2009). Among those students with disabilities included in the most recent National Assessment of Educational Progress (NAEP) study, a startling 83% of 4th graders and 92% of 8th graders were below grade level proficiency in mathematics (National Center for Education Statistics, 2013). Because disability is dichotomized in these studies (i.e., students "have" or do not "have" a disability), identifying the specific national achievement profiles of students with ID at less severe levels is not possible.

Developmental research has shed more light on the language functioning of children with ID. For children with ID, language functioning (skills in the areas of syntax, morphology, expressive and receptive vocabulary knowledge, semantics, and pragmatics; Bloom & Lahey, 1978) is often a significant impairment for overall functioning. Most individuals with ID have receptive and expressive language delays (and debatably deficiencies) that go beyond what can be explained by mental age or the level of general cognitive functioning alone (Miller et al., 1981; Rondal, 2003; Rosenberg & Abbeduto, 1993). 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).

When assessing mathematics ability, differences in item modality affect performance for children who are language minorities and children who experience language difficulties/disabilities (Abedi et al., 2001; Abedi & Lord, 2001; Abedi et al., 1997, 1998; Shaftel et al., 2006). Paper/pencil or verbal formats are common for mathematics assessments, while less linguistically demanding formats are more rare (e.g., using manipulatives,

DOI: 10.1352/1944-7558-120.6.542

pictorial displays, or pointing/gesturing formats; Parmar et al., 1996).

Walker and Arnault (1991) suspected that test format issues influence the factor structure of the KM-R. The KM-R is among several measures of mathematics ability that rely heavily on language as the primary modality of question delivery and response delivery. Linguistic features such as abstract or ambiguous language, unfamiliar vocabulary words, passive verb use, long nominal phrases, use of conditional clauses, and open response rather than multiple choice answer formats have all been implicated as linguistic features that make mathematics items more difficult (Kopriva, 1999; Shaftel et al., 2006). Many of the KM-R items involve one or more of these features of linguistic complexity. For example, Numeration item 2, "Hold up as many fingers as there are sheep in this picture," in addition to being an open response format, is an imperative sentence with the object and crucial pieces of the instructions buried in a correlative conjunction phrase, which also contains a prepositional phrase. Addition item 2, "Five baseballs and two soccer balls are how many balls in all," involves an object-verb-subject sentence order, with a compound direct object in which both nouns are modified by adjectives, and the subject (balls) is sandwiched between "how many" (an interrogative modifier) and "in all" (a prepositional phrase). Items with these complex linguistic features are prevalent across the KM-R.

The extent to which the language-heavy format influences the factor structure of the KM-R has not been investigated, nor has the discriminant validity of the KM-R been compared to tests of language skill. These concerns about the construct and content validity of the KM-R may be especially relevant for children with ID and associated language difficulties. KM-R mathematics items that involve complex syntax, morphology, vocabulary, and semantics may be more difficult for children with ID to answer.

This study addresses the extent to which a language-heavy assessment like the KM-R may unintentionally become a measure of *language* ability, as opposed to a measure of *mathematics* ability, when used with children with ID and associated language difficulties. Three rival hypotheses of the roles of *mathematics* and *language* are considered (presented in Figure 2): (A) that performance on math items is predicted by an ability for mathematics which is separate but perhaps related to language ability (i.e., the KM-R is a unidimensional test of *mathematics* for this population which demonstrates discriminant validity with measures of language ability; (B) that mathematics item performance is solely the result of language abilities (i.e., the KM-R demonstrates no discriminant validity with measures of language ability, but instead is a unidimensional test of language for this population), or (C) that math item performance is predicted by both language ability as well as a separate but additional ability to solve mathematics conceptual and computational problems (i.e., the KM-R does not demonstrate discriminant validity with measures of language ability; it is a multidimensional, mixed test of language and mathematics for this population). Figure 2 illustrates *language* and *mathematics* as latent constructs in ellipses. For the *language* factor, subtest and test indicators are depicted as rectangles. For the *mathematics* factor, items are depicted as rectangles. Each of these hypotheses was tested in a series of confirmatory models of item performance. This investigation of the factor structure of the KM-R adds to the body of literature characterizing the construct validity of this popular mathematics assessment, especially for a population of children with ID at less severe levels and language difficulties.

Method

The participants of this study were drawn from a 5-year, longitudinal reading intervention study designed to test the efficacy of reading programs for students with ID at less severe levels (Sevcik, 2005). Participants were selected for the parent study 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), who met school district criteria for ID at less severe levels, and who were eligible for special education services for children with ID. All school districts in the current study based diagnoses of ID at less severe levels on Georgia Department of Education (GADOE, 2011) standards, which specify the following diagnostic description for ID at less severe levels:

> A mild intellectual disability is defined by the GADOE as intellectual functioning ranging

DOI: 10.1352/1944-7558-120.6.542

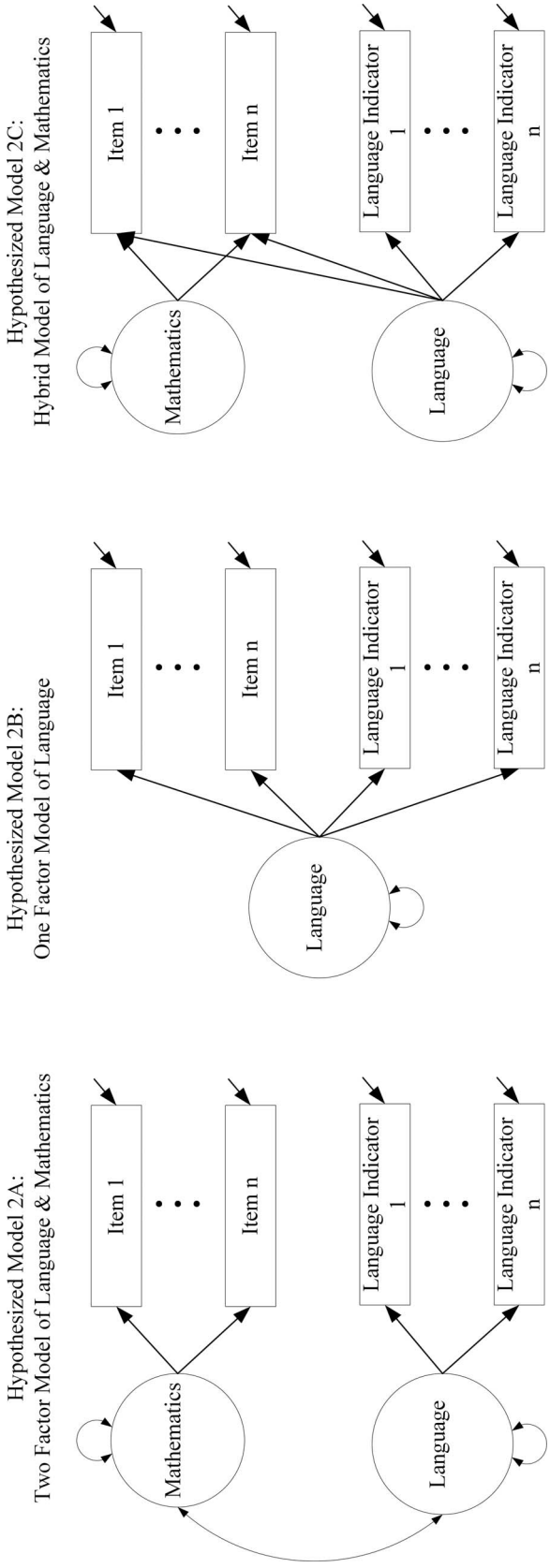

Figure 2. Schematic diagram for hypothesized models of language and mathematics. These are conceptual structural equation diagrams for confirmatory factor models of language versus math factors. Mean structure, residual variances, and link functions for dichotomous items are not shown.

DOI: 10.1352/1944-7558-120.6.542

Table 1
Descriptive Statistics for Demographic Variables

	N	Mean (*SD*)	Min – Max
Age (months)	264	111.25 (16.06)	80–147
PPVT Lang. Age	264	4.75 (1.65)	1.09–11.04
IQ	209	62.90 (9.48)	37–87
Grade level	264	3.36 (1.13)	2–5
Mother years of education	241	12.73 (3.02)	0–19
Father years of education	166	12.62 (3.59)	0–22

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

> between an upper IQ limit of approximately 70 to a lower IQ limit of approximately 55; and deficits in adaptive behavior that significantly limit a child's effectiveness in meeting the standards of maturation, learning, personal independence or social responsibility, and especially school performance that is expected of the individual's age level and cultural group. (Georgia Department of Education, 2011)

Consent packets were sent home for parents to review, and participation was allowed for those students who returned completed consent forms and who assented to participation at the time of baseline testing. Students were eligible for inclusion in the reading intervention study if they demonstrated difficulty in developing reading skills as determined by preliminary screening of reading performance on several standardized assessments of reading achievement (e.g., the Woodcock Reading Mastery Test; Woodcock, 1998). 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 or psychiatric disturbance, based on school records. Difficulty with oral language and articulation was not an exclusionary criterion for participants and every effort was made to include orally approximated correct responses for participants who struggled with articulation. Recruitment also attempted to balance the sample across the sexes.

Participants

A final sample of 264 children was selected for the current study from the baseline time point of the reading intervention study, representing fall of the academic year. Table 1 presents descriptive statistics for variables for the characteristics of the sample. This sample ranged in age from 80 months to 147 months, with a mean age of 111.25 months ($SD = 16.06$). The overall sample grade level mean was 3.36 ($SD = 1.13$). Approximately 64% of the sample was male ($n = 168$). The sample was racially and ethnically diverse (56% African American, 20% Caucasian, 16% Hispanic, 2% Asian, and 4% Multiracial). The mean level of education for mothers ($n = 241$ respondents) was 12.73 years ($SD = 3.02$), and the mean level of education for fathers ($n = 166$ respondents) was 12.62 years ($SD = 3.59$).

Because eligibility for a diagnosis of ID at less severe levels was determined by participants' local school districts, the range of IQ scores reported within students' individualized education programs (IEPs) was beyond the range defined by ID at these levels (IQs of between 50 and 70); however, these participants were considered representative of a population of children labeled as having ID at less severe levels and receiving special education services for students with ID within public school settings. The mean Peabody Picture Vocabulary Test (PPVT) language age of this sample was 4.75 years ($SD = 1.65$), and the mean IQ of this sample was 62.90 ($SD = 9.48$). Valid IQ scores were provided by participating schools for 209 of the total 264 students participating; these IQ assessments were not part of the current study's assessment battery. The participating schools used a variety of assessment instruments (e.g., Kaufman Assessment Battery for Children [K-ABC]; Differential Ability Scales [DAS]; Stanford-Binet Intelligence Scales; Universal Nonverbal Intelligence Test [Unit]; Wechsler Abbreviated Scale of Intelligence [WASI]; and Wechsler Intelligence Scale for Children [WISC], which were adminis-

DOI: 10.1352/1944-7558-120.6.542

tered onsite by affiliated, trained professionals (e.g., school psychologists, educational testing specialists, educational psychologists). Despite the fact that the IQ range of this school-based sample fell outside of the range of IQs typically accepted as ID at less severe levels, all of the students who participated in the current study were labeled as having ID at those levels by their schools. Missing IQ data occurred for 55 of 264 students across all 12 of the schools participating in the current study; however, all students in the current study (a) were diagnosed with ID by their school districts, (b) had current IEPs for their ID diagnoses, and (c) were receiving special education services at their schools, and were therefore identified as having ID at less severe levels for the purposes of inclusion in the current study.

Measures

Mathematics achievement measure. The current research study used the KeyMath-Revised Diagnostic Inventory (Connolly, 1988) Form A as a measure of mathematics achievement. Special education research indicates that elementary school-age children identified as having at less severe levels ID and receiving special education services often receive instruction in only basic mathematics skills like counting, numeration, quantity, and time and money skills (e.g., see Butler, Miller, Lee, & Pierce, 2001). Instructional emphasis on problem-solving skills and more advanced mathematical concepts may be limited for these students. Based on this trend in educational research and preliminary classroom observations by the current study team, six subscales of the KM-R (Form A) were selected as appropriate for administration to the participating students: Numeration, Geometry, Addition, Subtraction, Measurement, and Time and Money subscales. Seven of the KM-R subscales (Rational Numbers, Multiplication, Division, Mental Computation, Estimation, Interpreting Data, and Problem Solving) were deemed inappropriate given these students' limited mathematics instructional experiences.

Questions on the KM-R were administered orally with minimal visual support from an illustration array, and student responses were provided orally. For example, Numeration item 1 would involve showing an examinee a picture display board image of three sheep grazing in a pasture, asking the examinee, "How many sheep are in this picture," and recording the examinee's response as either correct or incorrect. The KM-R assessment was not timed. Each subscale was administered until students reached a ceiling with three consecutive incorrect responses. Correct responses to each item were recorded as 1 and incorrect responses were recorded as zerp. Table 2 contains percentages of the sample answering correctly as well as item-test correlations for the KM-R items used in subsequent analyses.

Split-half reliability coefficients for the KM-R assessment are reported in the technical manual for each subtest and grade level. Students in grades 1 through 5 of the normative sample all demonstrated reliability coefficients at or above .75 for the numeration subtest, at or above .72 for the geometry subtest, at or above .56 for the addition subtest, at or above .68 for the subtraction subtest, at or above .72 for the measurement subtest, and at or above .67 for the time and money subtest (Connolly, 1988). Model-based reliability in the form of R-squared coefficients will be presented for the current sample in the "Results" section.

Child language measures.

Vocabulary knowledge. Table 3 contains descriptive statistics for all language measures included in the study. 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, K. T., 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 ID at less severe levels. These assessments were administered such that basal scores and ceilings were established for all participants.

PPVT III. 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.

DOI: 10.1352/1944-7558-120.6.542

Table 2
Descriptive Statistics for KeyMath-Revised Items

Item	Percent Correct	Variance	Correlation With Total	Item	Percent Correct	Variance	Correlation With Total
Numeration 1	96.97	.03	.17	Subtraction 1	55.30	.25	.62
Numeration 2	78.41	.17	.40	Subtraction 2	2.27	.02	.09
Numeration 3	91.67	.08	.28	Subtraction 3	3.79	.04	.25
Numeration 4	32.95	.22	.58	Subtraction 4	28.79	.21	.66
Numeration 5	71.97	.20	.38	Subtraction 5	15.91	.13	.67
Numeration 6	36.74	.23	.64	Subtraction 6	5.68	.05	.50
Numeration 7	37.50	.24	.62	Subtraction 7	8.71	.08	.52
Numeration 8	15.53	.13	.73	Subtraction 8	1.14	.01	.24
Numeration 9	6.82	.06	.50	Subtraction 9	.38	.00	.22
Numeration 10	4.92	.05	.40	Subtraction 10	.38	.00	.22
Numeration 11	10.61	.10	.69	Subtraction 11	.38	.00	.22
Numeration 12	6.06	.06	.50	Subtraction 12	.00	.00	—
Numeration 13	5.30	.05	.55	Subtraction 13	.00	.00	—
Numeration 14	2.65	.03	.42	Subtraction 14	.00	.00	—
Numeration 15	.76	.01	.10	Subtraction 15	.00	.00	—
Numeration 16	.00	.00	—	Subtraction 16	.00	.00	—
Numeration 17	.38	.00	.11	Subtraction 17	.00	.00	—
Geometry 1	65.15	.23	.40	Subtraction 18	.00	.00	—
Geometry 2	34.09	.23	.52	Measurement 1	64.39	.23	.50
Geometry 3	61.74	.24	.51	Measurement 2	62.12	.24	.47
Geometry 4	15.53	.13	.51	Measurement 3	65.15	.23	.31
Geometry 5	48.86	.25	.41	Measurement 4	18.18	.15	.60
Geometry 6	42.05	.24	.62	Measurement 5	10.98	.10	.57
Geometry 7	35.98	.23	.50	Measurement 6	7.95	.07	.58
Geometry 8	26.89	.20	.54	Measurement 7	8.71	.08	.49
Geometry 9	12.88	.11	.37	Measurement 8	10.61	.10	.58
Geometry 10	9.47	.09	.35	Measurement 9	8.33	.08	.57
Geometry 11	6.06	.06	.28	Measurement 10	3.79	.04	.43
Geometry 12	4.92	.05	.40	Measurement 11	1.89	.02	.29
Geometry 13	.76	.01	.14	Measurement 12	.76	.01	.23
Geometry 14	3.79	.04	.29	Measurement 13	.38	.00	.09
Geometry 15	3.79	.04	.39	Measurement 14	.38	.00	.22
Geometry 16	.38	.00	.14	Measurement 15	.00	.00	—
Geometry 17	.00	.00	—	Measurement 16	.38	.00	.22
Geometry 18	.76	.01	.21	Measurement 17	.00	.00	—
Geometry 19	.00	.00	—	Measurement 18	.00	.00	—
Addition 1	76.89	.18	.37	Time & Money 1	44.32	.25	.45
Addition 2	66.67	.22	.41	Time & Money 2	56.82	.25	.41
Addition 3	16.29	.14	.43	Time & Money 3	39.02	.24	.44
Addition 4	41.67	.24	.57	Time & Money 4	16.67	.14	.57
Addition 5	32.95	.22	.62	Time & Money 5	18.18	.15	.69
Addition 6	17.42	.14	.66	Time & Money 6	21.97	.17	.62
Addition 7	20.45	.16	.63	Time & Money 7	6.82	.06	.47

(Table 2 continued)

DOI: 10.1352/1944-7558-120.6.542

Table 2
Continued

Item	Percent Correct	Variance	Correlation With Total	Item	Percent Correct	Variance	Correlation With Total
Addition 8	18.56	.15	.66	Time & Money 8	7.20	.07	.53
Addition 9	10.23	.09	.50	Time & Money 9	3.79	.04	.47
Addition 10	6.44	.06	.53	Time & Money 10	1.14	.01	.30
Addition 11	3.79	.04	.49	Time & Money 11	3.41	.03	.43
Addition 12	1.14	.01	.27	Time & Money 12	1.14	.01	.29
Addition 13	3.41	.03	.37	Time & Money 13	.38	.00	.21
Addition 14	.76	.01	.21	Time & Money 14	.38	.00	.22
Addition 15	.00	.00	—	Time & Money 15	.00	.00	—
Addition 16	.00	.00	—	Time & Money 16	.38	.00	.21
Addition 17	.00	.00	—	Time & Money 17	.00	.00	—
Addition 18	.00	.00	—	Time & Money 18	.00	.00	—

Note. Dashes indicate that correlations with the total score could not be calculated because none of the examinees provided correct responses. All other items, with the exception of Numeration items 15 and 17, Subtraction item 2, and Measurement item 13 significantly correlated with the total score at the $p < .05$ level.

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 demonstrates predictive validity when used with examinees who have special language profiles (e.g., children with ID scored over 20 standard points below children with typical language profiles on average) and correlates well with other measures of vocabulary and moderately well with measures of verbal ability (r = .66–.91 across various tests of oral language and verbal ability selected for comparison), indicating high construct validity (Dunn & Dunn, 1997). Total scores were used in the current analysis. Model-based reliability (R^2) for the current sample was .65.

EVT. The EVT was administered by presenting students with an illustration and asking them to name objects or actions, or to provide another word for the illustration. The assessment was not timed. Items were administered until students reached a ceiling of five consecutive incorrect responses.

The EVT demonstrates high reliability in both test-retest results and item uniformity in the

Table 3
Descriptive Statistics for Child Language Measures

Measure	CFD	WS	RS	FS	WC	SS	PPVT	EVT
Concepts & Directions (CFD)	1.00							
Word Structure (WS)	.69	1.00						
Recalling Sentences (RS)	.70	.73	1.00					
Formulating Sentences (FS)	.63	.70	.71	1.00				
Word Choices (WC)	.67	.63	.56	.61	1.00			
Sentence Structure (SS)	.72	.62	.57	.62	.70	1.00		
PPVT	.65	.67	.52	.61	.64	.68	1.00	
EVT	.65	.68	.60	.65	.63	.64	.69	1.00
Mean	12.79	10.69	17.61	11.48	20.97	14.18	67.25	49.46
SD	9.09	6.64	14.92	10.51	10.86	5.18	21.96	10.46

Note. All correlations are significant at the $p < .01$ level. CELF = *Clinical Evaluation of Language Fundamentals* (4th ed.) (Semel, Wiig, & Secord, 2003). PPVT = Peabody Picture Vocabulary Test III Form A (Dunn & Dunn, 1997). EVT = Expressive Vocabulary Test (Williams, K. T., 1997).

DOI: 10.1352/1944-7558-120.6.542

normative sample. The EVT also demonstrates high construct validity as evidenced by word frequency data, age differentiation, predictive validity when used with children who have special language profiles (e.g., children with ID scored over 30 standard points below children with typical language profiles on average), and correlation with other language measures requiring expression ($r = .47–.86$ across various tests of oral language and verbal ability selected for comparison). Total scores were used in the current analysis. Split half reliability has been reported as .91 (Williams, K. T., 1997). Model-based reliability (R^2) for the current sample was .64.

Syntactic and morphological functioning. Syntactic and morphological functioning can be conceptually defined as awareness of grammaticality, the rule-governed structure of language. The CELF-4 (*Clinical Evaluation of Language Fundamentals,* 4th ed.[Semel, Wiig, & Secord, 2003]) 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). Factor analytic studies of the CELF-4 support the measurement of one general language factor across subtests for an age range of 5 to 21 years, and high sensitivity and specificity for identifying language and learning disorders (at a cut score of 1.5 standard deviations below the mean performance of students with typical language profiles, the CELF-4 demonstrated a sensitivity of 1.00 and a specificity of .89 for children with language or learning disorders; Semel, Wiig, & Secord, 2003). For this sample of children, with average language age 4.80 years ($SD = 1.63$), the CELF-4 Language Structure Index subtests appropriate for children ages 5 to 8 years were used to measure children's syntactic and morphological functioning (the Word Structure subtest, Recalling Sentences subtest, Formulated Sentences subtest, and Sentence Structure subtest comprise the CELF-4 Language Structure Index score). For children identified as having ID, the Language Structure subtests of the CELF-4 all demonstrated reliabilities at and above .85 across content, time, and scorer (Semel, Wiig, & Secord, 2003).

Word structure. The Word Structure subtest of the CELF-4 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. Total scores were used in the current analysis. Model-based reliability (R^2) for the current sample was .65.

Recalling sentences. 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. The assessment was untimed. Ceiling was reached when students answered five consecutive items with four or more errors in repetition. Total scores were used in the current analysis. Model-based reliability (R^2) for the current sample was .55.

Formulated sentences. 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.'"). The subtest was untimed and administered until a ceiling of five consecutive scores of zero were obtained. Total scores were used in the current analysis. Model-based reliability (R^2) for the current sample was .61.

Sentence structure. 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. Total scores were used in the current analysis. Model-based reliability (R^2) for the current sample was .64.

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 Lan-

DOI: 10.1352/1944-7558-120.6.542

guage 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 Classes I were selected as the subtests to be included as indicators of the semantic aspects of child language profile. 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 across content, time, and scorer (Semel, Wiig, & Secord, 2003).

Concepts and following directions. 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. Total scores were used in the current analysis. Model-based reliability (R^2) for the current sample was .73.

Word classes I. The Word Classes 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. Total scores were used in the current analysis. Model-based reliability (R^2) for the current sample was .63.

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. 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. For the purposes of this study, the KM-R was administered to students before the PPVT, EVT, and CELF; however, each of these baseline measures was administered within two weeks of each other. 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. Standardized administration and scoring procedures were used to score the KM-R, PPVT, EVT, and CELF; however, raw scores were used in subsequent analyses. Both observed total 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.

Results

Analysis Overview

Statistical and conceptual assumptions of confirmatory factor analysis (CFA) were considered prior to analysis (for an overview, see, e.g., Bentler & Chou, 1987). The analyses consisted of two steps. In the first step of analysis, *language* and *mathematics* were analyzed independently to confirm that they were defensible single-factor structures. Separate confirmatory factor analyses were conducted for a one-factor *mathematics* model 1A and a one-factor *language* model 1B, using Mplus 7 (Muthen & Muthen, 2012). The one-factor *mathematics* model (1A) was indicated by dichotomously scored math items, and thus, weighted least squars estimation with mean and variance correction (WLSMV in Mplus) was used. For the

DOI: 10.1352/1944-7558-120.6.542

continuous *language* measures, relations were visually inspected for linearity. As is common for the population of children with ID, some of the *language* indicators were skewed, so the *language* CFA (model 1B) was fit with both the WLSMV estimator and with a robust maximum likelihood estimator (MLR in Mplus). The WLSMV and MLR estimators are commonly used for CFA with dichotomous and nonnormal data. The model fit and resulting parameters were comparable across estimation methods; for the sake of consistency, only WLSMV model results are reported.

Next, the central research question of this study, which sought to examine the role of child *language* ability in predicting item level mathematics assessment performance on the KeyMath-Revised Diagnostic Inventory of Essential Mathematics (KM-R), was examined. A series of three models were tested, one for each of the three hypotheses: (a) that performance on math items is predicted by an ability for mathematics which is separate but perhaps related to language ability, (b) that mathematics item performance is solely the result of language abilities, and (c) that math item performance is predicted by both language ability as well as a separate but additional ability to solve mathematics problems. Each of the three hypothesized models was fit via robust weighted least squares estimation (WLSMV) using Mplus 7 (Muthen & Muthen, 2012).

Baseline Model 1A: Mathematics

Two baseline factor models provide the background for this study. Model 1A tests the extent to which the mathematics items measured a single ability for this group of students. Model 1B tests the extent to which the eight language indicators measured a unitary, underlying ability of language proficiency. Models 2A-2C test the three main research hypotheses.

For the purposes of model estimation, only KM-R items with demonstrated variance (i.e., items without severe ceiling effects, on which at least 1% of the participants were able to provide a correct answer) were included in the baseline CFA analysis of Model 1A (see Table 2). The 77 KM-R items which demonstrated variance were Numeration items 1–15, Geometry items 1–15 and 18, Addition items 1–14, Subtraction items 1–8, Measurement items 1–12, and Time and Money items 1–12. This model was fit in Mplus 7 via weighted least squares with mean and variance correction (WLSMV) for items scored correct or incorrect (Muthen & Muthen, 2012). The fit statistics indicated that this one factor model of *mathematics* was an approximate good fit for the data, (χ^2 (2849) = 3656.79, $p < .001$, root mean square error of approximation (RMSEA) = .03, comparative fit index (CFI) = .91; for a discussion of fit, see Marsh, Hau, & Grayson, 2005; Marsh, Hau, & Wen, 2004).

Completely standardized factor loadings ranged from –.08 to .96. Numeration items 1 and 3 were not high quality indicators for this one factor model (Numeration item 1 $\lambda = -.08$ (.13) and Numeration item 3 $\lambda = .15$ (.12), with residual variances equal to .99 and .98, respectively). These items demonstrated serious floor effects for this sample, with 97% of respondents correctly answering Numeration item 1 and 92% of respondents correctly answering Numeration item 3. The substantive interpretation of Model 1A of *mathematics* suggested that the items were functioning reasonably well in this sample, with the exception of Numeration items 1 and 3. With regard to content validity of the KM-R, these results are consistent with a unidimensional mathematics assessment. These results did not differ substantially from the subsequent models, and will be detailed later.

Baseline Model 1B: Language

The one factor language model 1B consisted of eight indicators, the PPVT-III, the EVT, and the following CELF subtests: Word Structure, Recalling Sentences, Formulated Sentences, Sentence Structure, Concepts and Following Directions, and Word Classes I. Local and approximate fit statistics indicated that this model was an approximate good fit for the data, (χ^2 (20) = 84.51, $p < .001$, *RMSEA* = .11, *CFI* = .95, respectively. Completely standardized factor loadings ranged from .78 to .84, and residual variances ranged from .30 to .39.

The Role of Language in Predicting Mathematics Achievement

Hypothesized Model 2A: Two-factor model of *language* and *mathematics*. The *mathematics* and *language* factors described in the previous CFA analyses were specified as two latent factors allowed to covary with IQ as an exogenous predictor of both mathematics and language (effectively making IQ a covariate). The model

DOI: 10.1352/1944-7558-120.6.542

was not a good fit for the data, (χ^2 (3567) = 5281.18, $p < .001$; $CFI = .80$; $RMSEA = .05$). IQ was a moderate, positive predictor of both *mathematics* and *language*, $\beta_1 = .39$, $B_1 = 0.05$, $SE = .01$, and $\beta_2 = .41$, $B_2 = 0.05$, $SE = .01$, respectively. However, a high correlation between the two latent factors, $r = .83$, $SE = .03$, indicated that *mathematics* and *language* did not demonstrate adequate discriminant validity.

Hypothesized Model 2B: One factor model of *language* only. The one factor model for *language* (indicated by the 77 KM-R *mathematics* items and 8 *language* indicators previously specified) with IQ as an exogenous predictor of *language* was not a good fit for the data, (χ^2 (3569) = 5298.09, $p < .001$; $CFI = .80$; $RMSEA = .05$). Completely standardized factor loadings ranged from –26 to .94. As observed in the *mathematics* CFA analyses, Numeration items 1 and 3 were not high quality indicators for this one factor model. IQ was again a moderate, positive predictor of language, $\beta_1 = .41$, $B_1 = 0.05$, $SE = .0$.

Model 2B is a restricted version of Model 2A in which the covariance between language and mathematics is perfect, and can be compared to model 2A with a Chi-square difference test. This restriction did not fit well, compared to Model 2A ($\Delta\chi^2$ (2) = 44.38, $p < .001$), and thus, Model 2B was rejected.

Hypothesized Model 2C: Hybrid Two Factor Model of *Language* and *Mathematics*. The alternate, hybrid model was specified with the *language* and *mathematics* CFA models, but instead of a single latent *math-language* correlation, performance on each KM-R item was predicted by both *language* ability and *mathematics* ability. Once again, IQ was modeled as an exogenous predictor of both *mathematics* and *language*, in effect, becoming a covariate. The model was a good fit for the data, (χ^2 (3491) = 3876.79, $p < .001$; $CFI = .96$; $RMSEA = .02$). IQ was not a significant predictor of *mathematics*, $\beta_1 = -.11$, $B_1 = -0.01$, $SE = .01$, but was a moderate, positive predictor of *language*, $\beta_2 = .42$, $B_2 = 0.05$, $SE = .01$, respectively. Results from this Model 2C are presented in Table 4.

Table 4 shows the parameter estimates from Model 2C in Figure 2. The top portion of Table 4 shows estimates for the eight indicators of *language*. The bottom portion shows estimates for the 77 mathematics items. From left to right, the columns of Table 4 list (a) the outcome indicators, (b) the mean structure: regression intercepts for the language indicators and thresholds for the mathematics items, (c) the *language* factor loadings, (d) the *mathematics* factor loadings, (e) each indicator's residual variance (variance not accounted for by Model 2C), and (f) each indicator's R^2 (model implied reliability). Both completely standardized estimates and unstandardized estimates with standard errors are included.

Although Numeration items 4, 6, and 7, Geometry items 2, 7, and 8, Addition items 4 and 5, Subtraction item 4, and Time and Money items 1 and 3 were significant indicators of mathematics in Model 1A (the single factor CFA of all 77 KM-R items), they were no longer significant indicators of *mathematics* in this hybrid model. Several items evidenced a negative pattern of factor loading, indicating that respondents with higher *mathematics* ability would be less likely to answer them correctly (Numeration items 1–3 and 5, Geometry items 1, 3, 5, and 6, Addition items 1 and 2, Subtraction item 1, Measurement items 1, 2, and 3, and Time and Money item 2). Other items were still significant indicators of *mathematics*; however, their factor loadings were low enough to indicate that they were no longer quality indicators of *mathematics* (Numeration items 8 and 11; Geometry items 4, 9, and 10; Addition items 3, 6, 7, 8, and 9; Subtraction items 5 and 7; Measurement items 4, 5, and 8; Time and Money items 4, 5, and 6). It should be noted that the items that were no longer significant or salient indicators of *mathematics* consistently appeared at the beginning of each subtest.

Similarly, several KM-R items, usually appearing toward the ends of each subtest, demonstrated low factor loadings on *language* (Numeration items 1 and 15; Geometry items 13 and 18; Addition items 12 and 14; Time and Money item 12). Subtraction item 2 was the only KM-R item which was not a significant indicator of language, $\lambda = .08$ (.08).

Summary of hypothesized model testing. Table 5 displays the overall fit of the five models tested in these analyses. The two preliminary models, 1A and 1B, show that individually, the language tests indicated a coherent latent factor of *language* ability, and the KM-R items indicated a coherent latent ability as well. In the sequence of structural models 2A-C, Model 2A was not a good fit for the data and had an exceedingly high correlation (r = .83) between the *language* and *mathematics* factors, suggesting that the KM-R

DOI: 10.1352/1944-7558-120.6.542

Table 4
Estimates for Model 2C: Hybrid Dual Effects Model for Mathematics and Language

	Intercept/Threshold		Language Factor Loadings		Math Factor Loadings			
Indicator	STD	UnSTD (*SE*)	STD	UnSTD (*SE*)	STD	UnSTD (*SE*)	Residual Variance	R^2
CFD	– 1.16	–10.29 (4.59)	.75	6.01 (.53)	—	—	.44	.56
WS	– .46	–2.86 (2.79)	.69	3.85 (.41)	—	—	.53	.47
RS	– .50	–6.98 (7.29)	.62	7.77 (.83)	—	—	.62	.38
FS	– .98	–10.07 (4.89)	.66	6.21 (.65)	—	—	.56	.44
WC	– 1.14	–11.85 (5.33)	.80	7.56 (.75)	—	—	.36	.64
SS	.61	3.19 (2.55)	.80	3.79 (.36)	—	—	.36	.64
PPVT	1.24	27.60 (9.84)	.78	15.88 (1.31)	—	—	.39	.61
EVT	2.52	25.35 (4.82)	.70	6.43 (.62)	—	—	.51	.49
NUM1	– .81	– .83 (1.83)	.25	.23 (.11)	– .83	– .84 (.07)	.25	.76
NUM2	2.65	2.78 (.88)	.60	.58 (.07)	– .46	– .48 (.06)	.44	.60
NUM3	2.00	2.07 (1.25)	.43	.40 (.11)	– .69	– .71 (.07)	.34	.68
NUM4	2.90	3.04 (.66)	.71	.67 (.06)	.07	.07 (.04)	.55	.50
NUM5	1.64	1.69 (.62)	.49	.46 (.08)	– .31	– .32 (.06)	.69	.36
NUM6	2.92	3.10 (.64)	.80	.77 (.05)	.00	.00 (.05)	.41	.64
NUM7	1.89	2.01 (.64)	.80	.77 (.05)	– .01	– .01 (.05)	.40	.64
NUM8	2.16	2.33 (.80)	.95	.93 (.03)	.19	.20 (.05)	.10	.91
NUM9	3.66	3.84 (1.21)	.82	.78 (.06)	.38	.39 (.06)	.24	.78
NUM10	3.50	3.60 (1.03)	.68	.63 (.08)	.43	.44 (.06)	.41	.62
NUM11	3.05	3.26 (.89)	.91	.89 (.04)	.29	.31 (.05)	.12	.89
NUM12	3.05	3.19 (1.02)	.81	.77 (.06)	.43	.45 (.05)	.21	.81
NUM13	4.85	5.08 (1.32)	.83	.78 (.06)	.45	.47 (.06)	.17	.85
NUM14	4.73	4.87 (1.71)	.71	.67 (.07)	.59	.61 (.07)	.19	.82
NUM15	2.11	2.12 (1.51)	.28	.25 (.04)	.73	.73 (.04)	.40	.60
GEO1	.89	.92 (.56)	.55	.51 (.07)	– .24	– .25 (.06)	.67	.37
GEO2	2.00	2.06 (.60)	.61	.57 (.07)	.05	.05 (.05)	.67	.37
GEO3	2.21	2.35 (.66)	.72	.69 (.07)	– .29	– .30 (.05)	.43	.62
GEO4	3.21	3.33 (.85)	.69	.65 (.07)	.22	.23 (.05)	.53	.51
GEO5	1.77	1.83 (.59)	.58	.54 (.06)	– .14	– .14 (.05)	.69	.36
GEO6	2.88	3.07 (.62)	.80	.77 (.05)	– .09	– .10 (.05)	.40	.65
GEO7	2.36	2.46 (.65)	.65	.61 (.06)	– .01	– .01 (.05)	.62	.42
GEO8	3.07	3.19 (.73)	.67	.63 (.07)	.09	.09 (.04)	.60	.45
GEO9	2.85	2.90 (.84)	.50	.46 (.09)	.27	.27 (.06)	.72	.30
GEO10	2.84	2.89 (.82)	.54	.50 (.10)	.33	.33 (.06)	.64	.38
GEO11	3.45	3.49 (1.07)	.51	.47 (.10)	.54	.54 (.06)	.49	.52
GEO12	3.31	3.41 (1.30)	.68	.64 (.07)	.45	.46 (.06)	.38	.64
GEO13	2.04	2.04 (34.00)	.25	.23 (.05)	.91	.91 (.07)	.13	.87
GEO14	3.88	3.93 (1.27)	.52	.47 (.11)	.61	.61 (.06)	.40	.61
GEO15	3.90	4.00 (1.66)	.67	.63 (.09)	.59	.60 (.05)	.25	.76
GEO18	6.35	6.35 (3.08)	.18	.17 (.07)	.94	.93 (.04)	.11	.89
ADD1	.35	.37 (.62)	.59	.56 (.07)	– .36	– .37 (.06)	.55	.50

(Table 4 continued)

DOI: 10.1352/1944-7558-120.6.542

Table 4
Continued

	Intercept/Threshold		Language Factor Loadings		Math Factor Loadings			
Indicator	STD	UnSTD (*SE*)	STD	UnSTD (*SE*)	STD	UnSTD (*SE*)	Residual Variance	R^2
ADD2	.93	.97 (.64)	.65	.61 (.06)	– .25	– .26 (.06)	.55	.50
ADD3	3.11	3.17 (.82)	.54	.50 (.09)	.22	.22 (.05)	.70	.33
ADD4	1.90	1.99 (.61)	.70	.67 (.06)	– .03	– .03 (.05)	.56	.49
ADD5	1.95	2.09 (.66)	.84	.82 (.04)	.01	.01 (.05)	.33	.71
ADD6	1.83	1.95 (.84)	.88	.85 (.05)	.17	.18 (.04)	.25	.79
ADD7	2.11	2.25 (.71)	.86	.83 (.04)	.15	.15 (.05)	.28	.75
ADD8	1.98	2.13 (.73)	.91	.88 (.04)	.18	.19 (.05)	.18	.84
ADD9	2.09	2.22 (.93)	.85	.81 (.05)	.28	.30 (.05)	.25	.78
ADD10	2.80	2.94 (1.00)	.84	.80 (.06)	.44	.45 (.06)	.16	.86
ADD11	3.15	3.26 (1.82)	.77	.72 (.08)	.54	.56 (.07)	.17	.85
ADD12	5.25	5.25 (2.13)	.27	.25 (.05)	.88	.88 (.02)	.17	.83
ADD13	2.72	2.80 (1.04)	.69	.64 (.07)	.50	.51 (.07)	.33	.69
ADD14	2.90	2.90 (31.63)	.22	.20 (.04)	.95	.94 (.07)	.07	.93
SUB1	3.19	3.43 (.72)	.82	.80 (.04)	– .23	– .24 (.06)	.31	.74
SUB2	2.68	2.68 (1.57)	.09	.08 (.08)	.71	.70 (.07)	.50	.50
SUB3	2.79	2.82 (.82)	.52	.48 (.13)	.54	.55 (.08)	.47	.54
SUB4	4.39	4.67 (.75)	.84	.81 (.04)	.08	.09 (.05)	.34	.70
SUB5	4.02	4.29 (.84)	.90	.87 (.04)	.24	.25 (.05)	.19	.84
SUB6	3.80	3.96 (.04)	.78	.73 (.08)	.40	.42 (.06)	.29	.74
SUB7	2.63	2.76 (.88)	.79	.75 (.06)	.29	.30 (.06)	.34	.69
SUB8	2.22	2.23 (9.38)	.42	.38 (.08)	.77	.77 (.03)	.26	.74
ME1	2.40	2.53 (.58)	.68	.65 (.06)	– .28	– .30 (.06)	.50	.56
ME2	1.71	1.80 (.61)	.68	.65 (.06)	– .25	– .26 (.05)	.51	.54
ME3	1.68	1.73 (.70)	.47	.44 (.08)	– .24	– .25 (.05)	.75	.29
ME4	3.21	3.37 (.74)	.79	.75 (.06)	.20	.21 (.05)	.40	.64
ME5	3.20	3.39 (.87)	.85	.82 (.05)	.24	.25 (.05)	.27	.76
ME6	5.12	5.38 (1.33)	.84	.80 (.06)	.42	.43 (.06)	.17	.85
ME7	3.32	3.44 (.99)	.70	.66 (.08)	.34	.35 (.06)	.45	.58
ME8	3.52	3.70 (.97)	.80	.76 (.06)	.29	.30 (.05)	.33	.70
ME9	3.62	3.81 (1.08)	.83	.79 (.06)	.36	.38 (.06)	.24	.79
ME10	5.01	5.17 (1.53)	.72	.67 (.08)	.50	.51 (.08)	.29	.73
ME11	4.37	4.42 (1.26)	.52	.48 (.05)	.61	.61 (.07)	.40	.61
ME12	3.92	3.93 (9.37)	.39	.36 (.06)	.83	.83 (.05)	.19	.81
TM1	1.11	1.15 (.61)	.61	.57 (.07)	– .07	– .07 (.04)	.67	.38
TM2	1.23	1.27 (.60)	.55	.51 (.07)	– .13	– .14 (.04)	.72	.33
TM3	2.49	2.55 (.65)	.54	.51 (.07)	.06	.06 (.05)	.74	.30
TM4	3.44	3.62 (.79)	.79	.75 (.06)	.16	.17 (.05)	.41	.63
TM5	2.98	3.19 (.78)	.90	.87 (.04)	.19	.21 (.05)	.20	.83
TM6	2.45	2.59 (.69)	.81	.77 (.05)	.16	.17 (.05)	.38	.66
TM7	2.73	2.87 (1.19)	.82	.79 (.06)	.34	.36 (.06)	.26	.77
TM8	2.00	2.09 (.92)	.78	.74 (.07)	.36	.37 (.06)	.32	.71

(Table 4 continued)

DOI: 10.1352/1944-7558-120.6.542

Table 4
Continued

	Intercept/Threshold		Language Factor Loadings		Math Factor Loadings			
Indicator	STD	UnSTD (*SE*)	STD	UnSTD (*SE*)	STD	UnSTD (*SE*)	Residual Variance	R^2
TM9	2.65	2.77 (1.05)	.81	.77 (.07)	.53	.55 (.07)	.12	.89
TM10	3.08	3.10 (3.76)	.47	.43 (.07)	.81	.81 (.03)	.16	.84
TM11	4.19	4.33 (1.45)	.74	.70 (.09)	.57	.59 (.07)	.17	.84
TM12	2.04	2.04 (34.00)	.25	.23 (.05)	.92	.91 (.05)	.12	.89

Note. Dashes indicate a parameter not estimated because of the way the model was defined (see Figure 1). Model fit: $\chi^2(3408) = 3844.42$, $p < .001$, *RMSEA* = .02, *CFI* = .96. See text for details.
STD and UnSTD = standardized and unstandardized model estimates. Under Indicators: Clinical Evaluation of Language Fundamentals (CELF) subtests = Concepts and Following Directions (CFD), Word Structure (WS), Recalling Sentences (RS), Formulated Sentences (FS), Word Classes I (WC), and Sentence Structure (SS). PPVT = Peabody Picture Vocabulary Test III; .EVT = Expressive Vocabulary Test. KeyMath-Revised subtests = Numeration (NUM), Geometry (GEO), Addition (ADD), Subtraction (SUB), Measurement (ME), and Time and Money (TM). Thus, NUM# indicates subtest and item number. For example, NUM1 indicates Numeration item 1.

items could lack discriminant validity with measures of language. The fit for Model 2B was worse than the inadequate fit demonstrated by Model 2A, indicating that although the KM-R items did not have discriminant validity with measures of language, their variance could not be entirely explained by *language* ability. Hybrid Model 2C demonstrated approximate good fit for the data and provided the most reasonable structure for the 8 language tests and KM-R items

Table 5
Fit Statistics for All Models Tested

Initial Measurement Models		χ^2	*df*	*p*	*CFI*	*RMSEA*	Note
1A	Baseline Mathematics CFA	3656.79	2849	<.001	.91	.03	
1B	Baseline Language CFA	84.51	20	<.001	.95	.11	Fit with MLR was comparable to fit with WLSMV
Hypothesized Structural Models		χ^2	*df*	*p*	*CFI*	*RMSEA*	Note
2A	2 Factors: Language & Math	5281.18	3567	<.001	.80	.05	$R = .83$ $\beta_{IQ–Math} = .39$ $\beta_{IQ–Lang} = .41$
2B	Language only	5298.09	3569	<.001	.80	.05	$\chi^2\ (2) = 44.38$, $p < .01$ $\beta_{IQ–Lang} = .41$
2C	Dual Effects of Language & Math	3876.79	3491	<.001	.96	.02	$\beta_{IQ–Math} = –.11$ *NS* $\beta_{IQ–Lang} = .42$

Note. Models 1A and 1B are presented as reference points for the basic validity of the separate measures. Model 2A had a correlation of .87 between the *Language*, and *Mathematics* factors. Model 2B significantly degraded fit of Model 2A as a baseline model. RMSEA = root mean square error of approximation; CFI = comparative fit index. MLR = maximum likelihood estimator; and WLSMV = Weighted least squars estimation with mean and variance correction, both in Mplus 7 (Muthen & Muthen, 2012).

DOI: 10.1352/1944-7558-120.6.542

in this sample. The allowance for *language* ability to predict KM-R item performance shed light on the exceedingly high correlation between the *language* and *mathematics* factors in Model 2A; both *language* and *mathematics* abilities predicted KM-R item performance.

Discussion

The current study sought to examine the role of language in predicting item-level mathematics achievement among children with ID at less severe levels. Three hypothesized models were examined.

In Model 2A, (which suggested that that *language* and *mathematics* were two separate factors defined by their separate prediction of language and mathematics items respectively for this population) *mathematics* did not demonstrate adequate fit to the data. The inadequate model fit combined with the high correlation between *language* and *mathematics* as separate factors suggested that the KM-R was not a unidimensional measure of *mathematics* and did not demonstrate discriminant validity with *language*. Model 2B (which suggested that that because children with ID and associated language difficulties may have been unable to engage the KM-R item format and access the mathematical content of the items, *language* was the sole predictor of mathematics item performance) also failed to demonstrate adequate fit statistics and significantly degraded fit with Model 2A as a baseline, indicating that although the KM-R *mathematics* factor did not demonstrate discriminant validity with measures of *language*, language was not the unidimensional factor solely predicting performance on the KM-R. The rejection of Model 2B is consistent with CHC theory's assertion that language and mathematics are indeed separable constructs for this population of children with ID at less severe levels.

Hypothesized Model 2C (which suggested that *language* was a predictor of performance in language-heavy math items, but *mathematics* also retained some unique predictive validity for this population) provided the most reasonable measurement structure for *language* and *mathematics*. Model 2C, which allowed for KM-R items to be predicted by both *language* and *mathematics* demonstrated good fit to the data.

IQ was included in each model analysis as a covariate (i.e., exogenous predictor) of the latent factors under investigation. Given that this particular sample of children with ID evidenced relatively low IQ (*mean* = 62.90, *SD* = 9.48) and floor effects on both the *language* and *mathematics* assessments included in this study, one might expect that IQ could entirely explain poor performance on indicators for both factors. Indeed, the role of IQ can be examined in order to see if *general intelligence* provides a source of overlap for these measures that might make the KM-R appear to be multidimensional for this population when in fact, it is not. However, consistent with CHC theory, *general intelligence* displayed only a moderate, positive relationship with both *language* and *mathematics*. (Note that this relationship with *mathematics*, while evident in Model 2A, was no longer evident in Hybrid Model 2C due to the fact that KM-R items retained little variance unique to *mathematics* when *language* was allowed to cross-load in the model.) *Language*, *mathematics*, and *general intelligence* were indeed separate constructs for this population of test users with ID.

Some researchers or practitioners may be tempted to interpret findings that the KM-R is a multidimensional instrument for children with ID as evidence that *language* and *mathematics* are not separable domains within the cognitive profile of this population. However, our findings indicate that (a) the language domain is unique and complete within itself, (b) as evidenced by the poor fit of Model 2B, *language* and *mathematics* are not one domain for this population, and (c) IQ does not completely explain the overlap between *language* and *mathematics* in predicting students' performances on the KM-R. For the population of children with ID, *general intelligence*, *language*, and *mathematics* are separable constructs, and the KM-R is a measure of both *language* and *mathematics* abilities.

Validity of the KM-R

The KM-R items did not demonstrate discriminant validity with the *language* measures in this sample of children with ID. Rather, the KM-R items represent a mix of content, predicted differentially by the constructs of both *language* and *mathematics*. Items toward the beginnings of subtests tended to be predicted mostly by *language*, while items toward the ends of subtests tended to be predicted mostly by *mathematics*. Several of the KM-R items included in this study were entirely predicted by *language*, while only one

DOI: 10.1352/1944-7558-120.6.542

Table 6
Illustrating the KM-R Language Threshold

	Participant Characteristics			
	"Average Joe"	"Amy"	"Bea"	"Cecil"
Ability Profile	Average Language & Math	High Lang. Low Math	Low Lang. High Math	High Lang. High Math
Lang. Factor Score	0.00	1.00	–1.00	1.00
Math Factor Score	0.00	–1.00	1.00	1.00
	Model Predicted Geometry Subtest Performance			
Item				
Geo 1	0	1	0	1
Geo 2	0	1	0	1
Geo 3	0	1	0	1
Geo 4	0	1	0	1
Geo 5	0	1	0	1
Geo 6	0	1	0	1
Geo 7	0	1	0	1
Geo 8	0	1	0	1
Geo 9	0	1	0	1
Geo 10	0	1	0	1
Geo 11	0	0	1	1
Geo 12	0	1	0	1
Geo 13	0	0	1	1
Geo 14	0	0	1	1
Geo 15	0	1	0	1
Total Score	0	At least 12	0	At least 15

Note. KM-R = KeyMath-Revised Diagnostic Inventory of Essential Mathematics; Lang. = Language; Geo = Geometry.

of the KM-R items was predicted entirely by *mathematics* (see Table 4).

The multidimensionality of the KM-R affects interpretations of its content validity as well. Content validity, or the extent to which a particular set of items represents the domain or construct being measured by an assessment, is typically evaluated based on criteria defined by experts in a particular field (see, e.g., Crocker & Algina, 2008). However, because the KM-R items appear to be measures of both *language* and *mathematics* ability for this population of test users, the mathematics domain is not the only construct being assessed by this instrument. Discussions of content validity in the absence of construct validity are not possible, and the results of the current analysis highlight not only the effect of the language content of the KM-R items, but also the KM-R's limited potential for measuring math content when used with children with ID.

Language as a Facilitating, Potentially Overwhelming Factor

Given that the standardized testing procedures of the KM-R dictate that three, consecutive, incorrect responses establish a ceiling on all subtests, the fact that items toward the beginnings of subtests were largely predicted by examinee *language* ability is not trivial. In effect, *language* ability appeared to create a kind of threshold effect. In general, only examinees with *language* abilities high enough to succeed on items toward the beginnings of these subtests were able to proceed through the KM-R to access the items predicted largely by *mathematics* abilities toward the ends of subtests. For the purposes of demonstrating the *language* threshold effect, Table 6 compares the model-predicted item

DOI: 10.1352/1944-7558-120.6.542

responses on the Geometry subtest of the KM-R for four participants of varying levels of language and math ability.

The first 10 items on the Geometry subtest are best predicted by *language* ability (Table 4). In fact, all of the Geometry items included in the research model are all significant indicators of *language* ability. The first 10 Geometry items are not well explained by *mathematics* ability. Only Geometry items 11 and beyond begin to demonstrate significant and salient *mathematics* factor loadings.

The top portion of Table 6 lists the ability profile, language factor score, and mathematics factor score for each of four fictitious participants for whom we will imagine that we know the true *language* and *mathematics* abilities. The bottom portion of Table 6 lists Geometry items 1-15 with model predicted responses for each participant. From left to right, the participants are as follows: (a) Fictitious Participant "Average Joe," who represents the average *language* and *mathematics* abilities of the this sample of children with ID, (b) Fictitious Participant "Amy", who has a high *language* ability, a low *mathematics* ability, (c) Fictitious Participant "Bea," who has a low *language* ability, a high *mathematics* ability, and (d) Fictitious Participant "Cecil," who has high *language* and *mathematics* abilities.

Note that only the examinees with the higher *language* abilities ("Amy" and "Cecil") were able to avoid reaching a ceiling point before Geometry item 3, thus allowing them to access the Geometry items with the most variance explained by *mathematics* ability. "Average Joe" and "Bea" reached ceiling points at Geometry item 3 and were never provided the opportunity to answer the Geometry items 10 and beyond, which were largely predicted by *mathematics* ability. Thus, despite the fact that "Amy" had a relatively low *mathematics* ability, this examinee is predicted to achieve one of the highest total scores on the Geometry subtest. "Bea", who had *high mathematics* ability, is predicted to achieve a total score of zero on the Geometry subtest.

The consequence of assuming that the KM-R is a unidimensional test of *mathematics* ability for this population of children with ID is that two children with the same level of *mathematics* ability may achieve very different scores as a result of the contribution of their *language* abilities. For example, "Bea" and "Cecil" have equivalent *mathematics* ability; however, "Cecil" has high *language* ability, while "Bea" has low *language* ability. "Cecil" was able to correctly answer Geometry items 1–15 and achieve a total score of at least 15 because he did not reach a ceiling on the items included in this model (note that Geometry items 16, 17, and 19 and beyond had zero variability for this sample and were not included in the model). Although one might expect "Bea" to achieve a similar total score because "Bea" has equivalent *mathematics* ability, this examinee was unable to correctly answer Geometry items 1 - 3, reaching a ceiling at Geometry item 3 and achieving a total score of zero on the Geometry subscale.

Limitations and Suggestions for Future Research

Construct specification. The six KM-R *mathematics* subtests used in the current study were selected to reflect the educational experiences of elementary school-aged children with ID at less severe levels, receiving special education services. Only the subtests of Numeration, Geometry, Addition, Subtraction, Measurement, and Time and Money were administered to the study sample. The purpose of the current study was the investigation of the discriminant validity of the KM-R with tests of *language* abilities, and to that end, the factor structure of these six KM-R subtests was evaluated to establish a baseline measurement model. Confirmatory factor analysis indicated that these subtests could be treated as a single, *mathematics* factor. Although the factor validity of the KM-R was not a central focus of this research, future research should examine the additional subtests of the KM-R for factor validity with this population of test users.

The *language* factor used within the current study was broadly defined to include syntax, morphology, vocabulary, and semantics. Although the language development features considered within the current study are core to the *Gc* domain (Schneider & McGrew, 2012), other features of the *Gc* domain such as listening ability, pragmatics, and communication ability may also play a role in predicting the KM-R mathematics achievement of children with ID. These dimensions of the *Gc* domain should be considered in future research.

Although the current study sought to examine the discriminant validity of the KM-R with measures of *language* ability, other broad domains of cognition may also be informative for discriminant validity in future research. Domains related

DOI: 10.1352/1944-7558-120.6.542

to memory and processing speed (*Short-Term Memory*, *Gsm, Long-Term Storage and Retrieval*, *Glr*, and *Processing Speed*, *Gs*) were not included in the current study and may be of interest in examining the role of *language* ability (*Gc*) in predicting the performance of children with ID on the KM-R mathematics achievement test. The phonological loop, a component of verbal working memory which works to temporarily store verbal information, represents a significant deficit for children with ID 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 ID can store and process in working memory, while also retrieving stored information for processing from long-term memory, may be an important predictor of their ability to perform on language-heavy assessment items. Broad, domain-free cognitive capacities like short-term memory (*Gsm*), long-term retrieval (*Glr*), and processing speed (*Gs*) should be considered as possible confounds of the performance of children with ID on the KM-R mathematics test in future research.

Other methodologies for examining multidimensionality. Understanding the multidimensionality of the KM-R can also be approached by examining the linguistic and mathematical features of the items themselves. It is possible that different levels of linguistic complexity across items may have different effects with respect to children's ability to solve math items. Additional research could examine the relationship between item linguistic complexity and children's language skills with controls for an item's mathematical difficulty. This may be approached from an explanatory item response theory framework (see for example De Boeck & Wilson, 2004), in which both the mathematics conceptual difficulty and the linguistic demands of each item can be modeled as item-specific effects. However, because the KM-R was designed with the assumption that it was a unidimensional mathematics achievement test, the current study could not control for the dimension of language ability without extensive reformatting at the item-level. Practically speaking, this level of control could be accomplished with the design of a mathematics assessment that has items expressing similar mathematics content with varying amounts of language in item prompts. 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 children with ID) could be examined for differential item functioning. Separate norms could be considered.

Similarly, future investigations could consider moving beyond the dichotomous scoring system to include a polytomous scoring system in order to investigate differential effects of errors among children with ID. While the finding that children's language skills predict their KM-R mathematics achievement patterns (in addition to contributions from a separate *mathematics* factor) provides support for the idea that children with ID are making errors of understanding, it does not conclusively prove this point. Difficulties in providing the correct answers to mathematics items may be the result of any number of errors (e.g., difficulty understanding question demands, difficulty retrieving appropriate math facts, incorrect algorithm selection, computational error; Goodstein, Kahn, & Cawley, 1976). The dichotomous (right/wrong) scoring system does not allow for a specific characterization of potentially different errors contributing to incorrect answers. Additional research could be informative regarding the specific error patterns of this population. For example, responses reflecting specific types of misunderstanding could be coded for use in a nominal item response model (e.g., Bock, 1972).

Generalizability of the current study. Practical considerations of IQ scores as they relate to actual classroom placement decisions should be taken into account when generalizing these results. The study sample was drawn from a population of children who had been identified by their schools as having ID at less severe levels 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. It is likely that educational experiences at the classroom, school, and school system levels, as well as potentially different effects for other cognitive abilities across children, may contribute to mathematics achievement in informative ways.

DOI: 10.1352/1944-7558-120.6.542

Addressing the Needs of Children With ID and Language Impairments

Previous studies also have questioned the content validity of popular mathematics assessments like the KM-R 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 et al., 1996). For children with ID specifically (and most likely for children who have language difficulties in general), the results of the current study suggest that mathematics test revision should also include special considerations for the language demands of the mathematics assessment items.

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 language-heavy mathematics items. Each of these testing accommodations was allowed in the current study, and yet children's language skills still overwhelmingly predicted their mathematics performance on the items at the beginnings of each subtest examined. Reading aloud, repetition, extra time, and redirection do not necessarily change the amount of linguistic information that children are asked to store and manipulate to comply with testing demands. Test developers may need to address the language demands of items during test development, rather than relying on testing accommodations after the fact, if language-heavy assessments are to be used for mathematics curriculum recommendations with students with ID at less severe levels.

Implications for Language Formatting in Mathematics Assessments

Previous research concerns about KM-R test formatting and discriminant validity (Walker & Arnault, 1991; Williams et al., 2007) appear to be major limitations of this popular mathematics assessment for some populations of students. Several researchers (e.g., Miller et al., 1981; Rondal, 2003; Rosenberg & Abbeduto, 1993) indicated that language functioning often represents a significant impairment for overall functioning in children with intellectual disabilities, affecting many domains of functioning and achievement (e.g., educational achievement, interpersonal relationships, emotion regulation). The results of this study support the notion that, depending on test formatting, language may be a major predictor of functioning in the specific area of mathematics achievement testing for children with ID at less severe levels.

Practical Applications

Children with ID at less severe levels 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. Reliable and valid measures of mathematics achievement are a necessity for the design of effective interventions. The results of the current study suggest that the KM-R, a popular mathematics achievement assessment for this population of children, is not a unidimensional test of mathematics; the KM-R is also a test of language ability. Practitioners should use caution in interpreting the KM-R mathematics testing performances of children with ID (and this most likely extends to populations of children who experience difficulty with language in general). 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. Language-based intervention also may be needed in order for this population of children to be successful on mathematics assessments that are formatted with language-heavy items.

Conclusions

Though tests of mathematics may routinely employ word problems as a method of assessing both arithmetic and problem-solving competency, the use of language-formatted items may adversely affect populations of students with language or intellectual disabilities. Factor validity and discriminant validity with measures of language ability should be included in the examinations of mathematics achievement measures' construct validity, especially for mathematics achievement assessments that rely on language-formatted items as the primary modality of item delivery.

The current research suggests that at least for the KeyMath Revised test in this sample of students with ID, both mathematics and language

DOI: 10.1352/1944-7558-120.6.542

abilities were being assessed. These findings suggest that diagnosis and treatment of mathematics achievement difficulty may be complicated by linguistic contamination of mathematics assessment item formats. It is possible that student limitations in language ability may overwhelm their performance on what is desired to be a test of mathematics. Further research to understand the separate roles of language and mathematics abilities, especially among students with language or intellectual disabilities, is crucially needed to devise more effective instruction and intervention.

References

Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). *NAEP math performance and test accommodations: Interactions with student language background* (CSE Tech. Rep. No. 536). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14,* 219–234. http://dx.doi.org/10.1207/S15324818AME1403_2

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE Tech. Rep. No. 478). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C., & Plummer, J. (1997). *Final report of language background as a variable in NAEP mathematics performance* (CSE Tech. Rep. No. 429). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Allen, M., & Yen, W. (2002). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Science, 4,* 417–423. http://dx.doi.org/10.1016/S1364-6613(00)01538-2

Bentler, P. M., & Chou, C. (1987). Practical issues in structural modeling. *Sociological Methods & Research, 16,* 78–117. http://dx.doi.org/10.1177/0049124187016001004

Bergeron, R., & Floyd, R. G. (2006). Broad cognitive abilities of children with mental retardation: An analysis of group and individual profiles. *American Journal on Mental Retardation, 111,* 417–432. http://dx.doi.org/10.1352/0895-8017(2006)111[417:BCAOCW]2.0.CO;2

Bloom, L. & Lahey, M. (1978). Language development and language disorders. New York, NY: Wiley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29–51. http://dx.doi.org/10.1007/BF02291411

Butler, F. M., Miller, S. P., Lee, K., & Pierce, T. (2001). Teaching mathematics to students with mild-to-moderate mental retardation: A review of the literature. *Mental Retardation, 39,* 20–31. http://dx.doi.org/10.1352/0047-6765(2001)039<0020:TMTSWM>2.0.CO;2

Carroll, J. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press.

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

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

Crocker, L., & Algina, J. (2008). *Introduction to classical & modern test theory*. Mason, OH: Cengage Learning.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

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

Geary, D. C. (1993). Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin, 114,* 345–362. http://dx.doi.org/10.1037/0033-2909.114.2.345

Georgia Department of Education. (2011). Special education services: Intellectual disabilities. Retrieved from http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Special-Education-Services/Pages/Intellectual-Disabilities.aspx.

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

DOI: 10.1352/1944-7558-120.6.542

10, 61–70. http://dx.doi.org/10.1177/002246697601000108

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

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57,* 253–270. http://dx.doi.org/10.1037/h0023816

Kopriva, R. (1999). *Ensuring accuracy in testing for English language learners: A practical guide for assessment development*. Washington, DC: Council of Chief State School Officers.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22,* 1–38.

Lyon, G. R., Shaywitz, S. E., Shaywitz, B. A. (2003). Part I: Defining dyslexia, comorbidity, teachers' knowledge of language and reading: A definition of dyslexia. *Annals of Dyslexia, 53*, 1–14. http://dx.doi.org/ 10.1007/s11881-003-0001-9

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 275–340). Mahwah, NJ: Lawrence Erlbaum.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(3), 320–341. http://dx.doi.org/10.1207/s15328007sem1103_2

Mazzoco, M. M. M., & Myers, G. F. (2003). Complexities in identifying and defining mathematics learning disability in the primary school-age years. *Annals of Dyslexia, 53*, 218–253. http://dx.doi.org/10.1007/s11881-003-0011-7

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. (2012). *Mplus: Statistical Analysis with Latent Variables* (v.7). Los Angeles, CA: Authors.

Naglieri, J. A., & Goldstein, S. (Eds.) (2009). *Practitioner's guide to assessing intelligence and achievement*. Hoboken, NJ: Wiley.

National Center for Education Statistics. (2013). *A first look: 2013 Mathematics and reading national assessment of educational progress at grades 4 and 8. (NCES Report 2014-451)*. Washington, DC: Institute of Education Sciences, U.S. Department of Education.

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., Frazita, R., & Cawley, J. F. (1996). Mathematics assessment for students with mild disabilities: An exploration of content validity. *Learning Disability Quarterly, 19*, 127–136. http://dx.doi.org/10.2307/1511253

Rondal, J. A. (2003). Atypical language development in individuals with mental retardation: Theoretical implications. In L. Abbeduto (Ed.), *International review of research in mental retardation: Language and communication in mental retardation, Volume* 27 (pp. 281–308). San Diego, CA: Elsevier, Inc.

Rosenberg, S., & Abbeduto, L. (1993). *Language and communication in mental retardation: Development, processes, and interventions*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 99–144). New York, NY: The Guilford Press.

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). Washingtgon, DC: Institute of Educational Sciences, U.S. Department of Education.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*, 105–126. http://dx.doi.org/10.1207/s15326977ea1102_2

DOI: 10.1352/1944-7558-120.6.542

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.

Swanson, H. L., & Jerman, O. (2006). Math disabilities: A selective meta-analysis of the literature. *Review of Educational Research, 76*, 249–274. http://dx.doi.org/10.3102/00346543076002249

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. http://dx.doi.org/10.1111/j.1365-2788.2006.00863.x

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. http://dx.doi.org/10.1177/15345084070320020201

Woodcock, R. W. (1998). *Woodcock Reading Mastery Test-Revised*. Circle Pines, MN: American Guidance Service.

Received 3/3/2014, accepted 12/22/2014.

This research was supported by the U.S. Department of Education Institute of Educational Sciences (Grant H324K040007), Rose A. Sevcik, Principal Investigator. We would like to thank the children who participated in this research project. Additionally, we are thankful for the many teachers and administrators in the Atlanta area school districts who allocated time during the busy school day for the students to participate. This manuscript was prepared, in part, for fulfillment of the requirements of Masters of Arts degree for Katherine T. Rhodes.

Authors:
Katherine T. Rhodes, Georgia State University (now at Ohio University); **Lee Branum-Martin, Robin D. Morris, MaryAnn Romski,** and **Rose A. Sevcik,** Georgia State University, Atlanta, Georgia.

Correspondence should be addressed to Katherine T. Rhodes, Georgia State University, Department of Psychology, P.O. Box 5010, Atlanta GA 30302-5010 USA (e-mail: krhodes1@student.gsu.edu).