PRI Middle School Follow-Up Study: Measures

Direct Assessments/Surveys of Students
Beginning in spring 2014, re-consented students were directly assessed with a number of mathematics and executive function tasks in the spring of each follow-up study year. Assessments were conducted by trained members of the research staff and were done in two individual sessions lasting about 45 minutes each.

- Standardized General Math Assessments
  - Key Math. The KeyMath 3 Diagnostic Assessment (‘KeyMath’; Connolly, 2007) is a comprehensive, norm-referenced measure of essential mathematical concepts and skills. KeyMath content covers the full spectrum of math concepts and skills that are typically taught in kindergarten through ninth grade and can be used with individuals aged 4½ through 21 years who are functioning at these instructional levels. The instrument has three concept areas. We used three subscales out of the five subscales in the Basic Concepts area:
    - Numeration - The Numeration subtest measures an individual’s understanding of whole and rational numbers. It covers topics such as identifying, representing, comparing, and rounding one-, two-, and three-digit numbers as well as fractions, decimal values, and percentages. It also covers advanced numeration concepts such as exponents, scientific notation, and square roots.
    - Algebra - The Algebra subtest measures an individual’s understanding of pre-algebraic and algebraic concepts. It covers topics such as sorting, classifying, and ordering by a variety of attributes; recognizing and describing patterns and functions; working with number sentences, operational properties, variables, expressions, equations, proportions, and functions; and representing mathematical relationships.
    - Geometry - The Geometry subtest measures an individual’s ability to analyze, describe, compare, and classify two- and three-dimensional shapes. It also covers topics such as spatial relationships and reasoning, coordinates, symmetry, and geometric modeling.
  - Because we did not administer the full five subscales in this content area, we were unable to obtain standard scores (the manual does not provide standard scores for individual subtests, only complete content areas and test total). However, we did obtain raw scores, the measure’s age- and grade-
scaled scores (adjusted for the age or grade of the student at time of assessment), and associated age and grade equivalent scores. Both the age and grade scaled scores have a mean of 10 and a SD of 3 for each grade level and age group. The manual provides some descriptive categories for these scaled scores:

- <=4: Well below average
- 5-7: Below average
- 8-12: Average
- 13-15: Above average
- >=16: Well above average

○ Woodcock Johnson Achievement Battery III: Quantitative Concepts Subtest ('WJ QC'; Woodcock, McGrew, & Mather, 2001). Individually-administered, Quantitative Concepts has 2 parts and assesses students’ knowledge of mathematical concepts, symbols, and vocabulary, including numbers, shapes, and sequences; it measures aspects of quantitative math knowledge and recognition of patterns in a series of numbers. We obtained raw scores, W-scores (the measure’s rescaled score that is an equal-interval scale) and standard scores that are age adjusted.

○ Nonstandard Math Assessments
  ○ Pre-Algebra Task ('Functional Thinking'). Developed by Bethany Rittle-Johnson (items taken from Lee, Ng, Bull, Pe, & Ho, 2011), this task consists of 6 ‘tables’ in which the student has to fill in the missing Input number, Output number, and Rule. The maximum possible score is 18 (3 points per table). Two function rules are addition, 2 are multiplication, and 2 are multiply-then-add.

  ○ Student’s Feelings About Math (FAM). In the first year of the project, we adapted a student survey from Karabenick and Maehr’s (2007) MSP Motivation Assessment Program. Students were asked 10 questions about how much they liked math, how good they were at math, how much they feel like their math teacher cared about them, etc. Responses were on a 5-point scale. The results demonstrated little variability in student responses. We were concerned that students may have felt they couldn’t answer honestly. During the second project year, this measure was replaced with TIMSS (Trends in International and Mathematics Science Study) (Martin, Mullis, & Foy, 2008; Mullis, Martin, Foy, & Arora, 2012). Students were asked 26 4-point-scale questions (i.e., Agree a lot, Agree a little, Disagree a little, Disagree a lot) about how they feel about math, including how much they enjoy doing math, how well they think they are doing in math, etc. While the TIMSS asks students very similar questions (and a greater number compared to the 10 FAM questions), students were asked to circle their answers on paper rather than respond verbally. This process was adopted to provide students with a
greater sense of confidentiality as they responded. TIMSS also provides national and international comparison data.

- **Domain-General Cognitive Measures: Executive Function and Visuospatial Skill**
  - Working Memory (‘Backward Corsi Blocks’; Vandierendonck, Kemps, Fastame, & Szmalec, 2004). This task involves a student’s working memory. Different numbers of squares light up in a sequence, and the student must then tap the squares in the reverse order from which they lit up. The task consists of 16 total trials made up of 8 2-trial items. The sequence length of squares increases from 2 to 8 across the activity. The score of interest for this measure is the highest span reached, or the longest sequence length that the student was administered and got at least 1 correct (span with at least 50% accuracy).
  - Attention Shifting (‘Hearts and Flowers’; Davidson, Amso, Anderson, & Diamond, 2006). This task tests a student’s ability to use attention shifting and inhibitory control by using the number pad to select congruent or incongruent sides of the screen based on different stimuli + rule combinations. The task consists of 12 congruent trials, 12 incongruent trials, and 48 mixed trials. We calculated the percent accuracy and mean response time for congruent trials, incongruent trials, all fixed trials (congruent and incongruent), and mixed trials.

- **Math-Specific Cognitive Measures**
  - Symbolic Number Comparison (‘Numbers’; Roussell & Noël, 2007). During the first year, this task assessed students’ symbolic Approximate Number System (ANS) acuity through the simultaneous presentation of two single-digit numbers. During the second year, two double-digit numbers were presented. Students were required to select which of the two numbers was larger. Scores of interest for this task include the total percent accuracy (mean and slope across ratios of the two presented digits), response time for correct responses (mean and slope across ratios of the two presented digits), and a Performance score, which includes both response time and error rate.
  - Nonsymbolic Number Comparison (‘Dots’; Roussell & Noël, 2007). Similar to the task above, this task presents two arrays of dots simultaneously and requires the student to determine which side of the screen contains more dots. Scores of interest for this task include the total percent accuracy (mean and slope across ratios of the two presented arrays), response time for correct responses (mean and slope across ratios of the two presented arrays), and a Performance score, which includes both response time and error rate.
    - In the first year of the project, this task was administered to all students, and then a slightly modified task was given to roughly half of the students, starting in the middle of the assessment period. This was because the original task appeared to be fairly difficult for
students, and we wanted to ensure usable data. In subsequent years, only the modified task (‘Color Dots’) was administered. In the modified task, two arrays of dots were presented simultaneously, one on the left side of the screen containing yellow dots and one on the right side containing blue dots. In half of the trials, blue was the correct answer and in half, yellow was the correct answer. Color and response side were fully counterbalanced. Trials consisted of 1200ms stimulus presentation followed by 1800ms of a fixation cross. Students were required to select via button press whether the right or the left set contained more dots. Seven ratios were presented, ranging from .33 (5 vs. 15) to .9 (9 vs. 10). The number of dots in each stimulus ranged from 5 to 15. Each ratio was presented 10 times for a total of 70 trials. Ratios and order of presentation were modeled after Odic, Hock, and Halberda (2014). To control for the possibility that students might choose a strategy based on visual cues rather than number of dots, the following visual properties of dot sets were varied using a modified version of the MATLAB code recommended by Gebuis and Reynvoet (2011) to generate stimuli: convex hull (area extended by a stimulus), total surface area (aggregate value of dot surfaces), average dot diameter, and density (convex hull divided by total surface area). In approximately one quarter of the trials all four visual properties were congruent with greater numerosity (i.e., the greater number of dots had a greater convex hull, surface area, etc.). In another approximate quarter of the trials, all four visual properties were incongruent with greater numerosity. In the remaining trials, visual properties were mixed congruent and incongruent.

- For the students administered both tasks in the first project year, zero-order correlations between scores of interest on the two measures were as follows:
  - Percent Accuracy: .342
  - Mean Response Time: .800
  - Performance Score: .748
  - Accuracy Slope: .112
  - Response Time Slope: .053

- Nonsymbolic and Symbolic Number Comparison Task (‘Mapping’; Lyons, Ansari, & Beilock, 2012). This task was introduced in the second year of the project. Students were presented with a symbolic number and nonsymbolic numerosity (i.e., a set of dots) and were required to select which of the two simultaneously presented stimuli represents the larger quantity. Trials consisted of 1200ms stimulus presentation followed by 1800ms of a fixation cross. Students were required to decide via button press whether the right or the left set was larger. Seven ratios were presented, ranging from .33 (5 vs.
The number of dots in each stimulus ranged from 5 to 15. Each ratio was presented 10 times for a total of 70 trials. Ratios and order of presentation were modeled after Odic, Hock, and Halberda (2014). Following data collection in Spring 2015, the task was modified as follows:

- **Description:** Participants were presented with one numerical stimulus centered at the top of the screen (either groups of dots or Arabic digits) and two stimuli at the bottom to the right and left and asked to respond via button press which stimulus at the bottom matched the numerical value of the stimulus at the top. Stimuli at the bottom were the opposite format of the stimulus at the top. There were 84 trials of 7 ratios, presented 6 times each. Seven ratios were presented: .33 (5 vs 15), .86 (6 vs 7), .5 (5 vs 10), .88 (7 vs 8), .67 (6 vs 9), .9 (9 vs 10), .8 (8 vs 10). In the first 42 trials, Arabic digits were presented at the top and sets of dots were presented below. In the second 42 trials, dots sets were at the top and digits were presented below. Trials consisted of 1500ms stimulus presentation followed by 2000ms of fixation for a total trial time of 3500ms. All three stimuli were presented within grey circles on a black background. During fixation, the grey circles remained but were empty of stimuli. The same stimuli were used from the symbolic and nonsymbolic comparison tasks, but the dot sets were colored green to reinforce the transition to a new task.

- **Details:** 84 Trials (7 ratios, 6 trials each). Phase 1 (first 42 trials): see a numeral at top, choose between dot sets below; Phase 2 (second 42 trials): see a dot set at top; choose between numerals below. Avoided subitizing range (1 - 4). Stimuli presented for 1500ms; Fixation circles 200ms; Total trial time = 3500 ms. Ratios = .33 (5 vs 15), .86 (6 vs 7), .5 (5 vs 10), .88 (7 vs 8), .67 (6 vs 9), .9 (9 vs 10), .8 (8 vs 10)

- **Nonsymbolic Enumeration Fluency** ('Groupitizing'; Starkey, 2014; Starkey & McCandliss, 2014). This task has a practice block, as well as 2 primary test components.
  - Before the test trials began, students were shown 27 practice trials in which they were asked only to press the number on their keypad that matches the numeral on the screen; this was done to prime students for this method of responding.
  - Next, during the first test section, students were presented with collections of dots to be enumerated under instructions for speeded responses. Students were presented with 64 (in Year 1) or 78 (in Year 2) trials in which they were instructed to rapidly indicate (through pressing the corresponding number on number keypad) the exact number of dots presented simultaneously on a computer screen. Arrays ranged from 1 to 9 in set size. Dot size, density, perimeter, and surface area were varied so as not to provide accurate cues to
numerosity. Primary scores of interest from this measure are accuracy and response time (for correct responses only) means and slopes of median response times across set sizes for random presentations of 1-3 (an indicator of subitizing ability), and random presentations of 5-7 and grouped presentations of 5-7 (the contrast of the two indicates a student’s groupitizing skill).

- In the second part, students were presented with 24 (in Year 1) or 42 (in Year 2) screens that had 2 or 3 numerals on them, and the student had to press the corresponding number key indicating the sum of the numbers on the screen. Sums ranged from 5 to 9. Primary scores of interest from this portion of the measure are:

**YEAR 1 VARIABLES**

- **Number Matching 1-3 IES**
  - Mean RT across all 1-3 NUMBER MATCHING observations (from the practice block) that were answered accurately/% accuracy for all 1-3 observations

- **Number Matching 5-7 IES**
  - Mean RT across all 5-7 observations that were answered accurately/% accuracy for all 5-7 observations

- **Random 1-3 IES** (subitizing speed)
  - Mean RT across all 1-3 nonsymbolic RANDOM observations that were answered accurately/% accuracy for all 1-3 observations

- **Random 5-7 IES** (counting speed)
  - Mean RT across all 5-7 nonsymbolic RANDOM observations that were answered accurately/% accuracy for all 5-7 observations

- **Random 5-7 Slope** (counting rate)
  - Median RT (for correct responses only) for nonsymbolic RANDOM 5, median RT for 6, median RT for 7 – slope of RT across those 3

- **Grouped 5-7 IES** (groupitizing speed)
  - Mean RT across all 5-7 nonsymbolic GROUPED observations that were answered accurately/% accuracy for all 5-7 observations

- **Grouped 5-7 Slope** (groupitizing rate)
  - Median RT (for correct responses only) from nonsymbolic block GROUPED 5, median RT for 6, median RT for 7 – slope of RT across those 3

- **Symbolic 6-8 IES** (for 3-group trials only, pooled across SubMax conditions; symbolic speed)
- Mean RT across all 6-8 observations (from symbolic block) that were answered accurately/% accuracy for all 6-8 observations
- Symbolic 6-8 Slope (for 3-group trials only, pooled across SubMax conditions; symbolic rate)
  - Median RT (for correct responses only) from symbolic block for 6, median RT for 7, median RT for 8 – slope of RT across those 3
- Subitizing Level (see description below)

**YEAR 2 VARIABLES**

- Number Matching 1-3 IES
  - Mean RT across all 1-3 NUMBER MATCHING observations (from the practice block) that were answered accurately/% accuracy for all 1-3 observations
- Number Matching 6-8 IES
  - Mean RT across all 6-8 observations that were answered accurately/% accuracy for all 6-8 observations
- Random 1-3 IES (subitizing speed)
  - Mean RT across all 1-3 RANDOM observations (from 1ST nonsymbolic block) that were answered accurately/% accuracy for all 1-3 observations
- Random 6-8 IES (counting speed)
  - Mean RT across all 6-8 RANDOM observations (from 1ST nonsymbolic block) that were answered accurately/% accuracy for all 6-8 observations
- Random 6-8 Slope (counting rate)
  - Median RT (for correct responses only) FROM 1ST nonsymbolic block for 6, median RT for 7, median RT for 8 – slope of RT across those 3
- Grouped 6-8 IES (pooled across SubMax conditions; groupitizing speed)
  - Mean RT across all 6-8 GROUPED observations (from 2ND nonsymbolic block) that were answered accurately/% accuracy for all 6-8 observations
- Grouped 6-8 Slope (pooled across SubMax conditions; groupitizing rate)
  - Median RT (for correct responses only) from 2nd nonsymbolic block for 6, median RT for 7, median RT for 8 – slope of RT across those 3
- Symbolic 6-8 IES (pooled across SubMax conditions; symbolic speed)
- Mean RT across all 6-8 GROUPED observations (from symbolic block) that were answered accurately/ accuracy for all 6-8 observations
- Symbolic 6-8 Slope (pooled across SubMax conditions; symbolic rate)
  - Median RT (for correct responses only) from symbolic block for 6, median RT for 7, median RT for 8 – slope of RT across those 3
- Subitizing Level (see description below)

**Subitizing levels**
- Values for this variable range from 0 to 5, and correspond to the level of subitizing skill that each student has achieved. A subitizing reaction time threshold was calculated independently for each student, and subitizing level was determined by whether a student demonstrated adequate accuracy and subitized set sizes consistently:
  - **Level 0**: Did not accurately hit buttons for 1, 2, and 3 (less than 85% accuracy)
  - **Level 1**: Able to accurately hit buttons for 1-3, but not for greater quantities (less than 85% for 4-8)
  - **Level 2**: Able to accurately hit buttons for all set sizes, but did not consistently subitize 3 (50% or fewer trials were below threshold)
  - **Level 3**: Subitized 3 consistently (on more than 50% of trials), but did not consistently subitize 4
  - **Level 4**: Subitized 4 consistently, but did not consistently subitize 5
  - **Level 5**: Subitized 5 consistently
- (Students who did not demonstrate an ability to subitize 3 are in Levels 2 and below.)

**More detailed description:**
In order to determine subitizing level, we first calculated each individual’s average subitizing speed (the median of their RTs for set sizes 1-2, which are known to be in the subitizing range even for infants). We then calculated an average incrementing speed for each individual, or the approximate time it takes them to count each item when not subitizing. We did this by calculating the difference between average subitizing speed and median RT for 6 and then dividing by 5 (because if they see 6 dots, they incremented 5 times after the first dot). This result equaled the average incrementing time per dot. We
then repeated this for 7 and 8 (because 6-8 are known to be beyond the subitizing range even for adults). The mean of these three results was then calculated to obtain each individual’s mean incrementing speed.

Based on this incrementing speed, we then set a threshold for each individual, which equaled their subitizing speed plus their incrementing speed. This threshold served as our estimate of what the individual’s RT would be if incrementing (counting) were used. Therefore, if an observed RT was below the threshold, this meant the student subitized during this trial. If an observed RT was above the threshold, this meant the student did not subitize.

There is evidence that children may be able to subitize a quantity, but do not do it consistently (Chi & Klahr, 1975; Svenson & Sjöberg, 1983). Because the students in this sample have subitizing skills that are still developing, we calculated the percentage of trials for each of the remaining set sizes (3-6 dots) that could be considered subitized (i.e., for which the RT fell below threshold).

Students were categorized into levels of subitizing skill based on these percentages, taking into consideration their accuracy for very small quantities. In assigning these levels, we used an iterative terminating procedure in which students had to meet all previous levels before consideration for the next level (for example, a student had to meet criteria for Level 3 before being considered for Level 4).

State/District-Collected Achievement Data
With help from the district, we were able to obtain state standardized test scores and course grades for each individual student.

- Tennessee Comprehensive Assessment Program (TCAP). This group-administered state assessment is given annually, usually in late spring, to all students in 3rd through 8th grade. It is a multiple-choice test and includes 4 content areas: Reading/Language Arts, Mathematics, Science, and Social Studies. We collected scores for the first two areas listed. There are 4 types of scores available for each content area. None of these scores are appropriate for examining growth, as they are not scaled longitudinally.
  - Scale Scores
  - Proficiency Levels (cut-offs for these categories can change each year, but students are categorized as below basic, basic, proficient, or advanced)
Normal Curve Equivalent (NCE): This is a transformation of the year-specific percentile that has equal-interval properties so they can be added and subtracted.

Percentile Scores: These are based on the population of test taking students in TN. The percentiles are grade- and subject-specific.

There were some students who were given the TCAP-MAAS (Modified Academic Achievement Standards) in 2013-14 rather than the TCAP (an alternative assessment decided on by the students’ IEP team). The MAAS scores should not be analyzed alongside TCAP scores as they are scaled differently.

- End of Course Grades: We obtained numeric end-of-course grades for Math and Language Arts. If a student had an ELD (English Language Development) Language Arts course grade but no typical Language Arts grade, we used the ELD grade. Occasionally, a student had both a typical Language Arts course grade AND an ELD course grade; in those cases, we took the typical grade. If a student had multiple typical course grades from different schools, we chose the school that the student attended the most.

**Teacher Questionnaire**

In addition to direct assessments and achievement score collection, students’ math teachers were asked to complete an online measure in the winter/spring of each study year. In the first year of the project, this measure was adapted from ECLS-K 1998-99 Spring 2004 Fifth Grade Child-Level Questionnaire by project staff and administered to all math teachers in an online survey. This survey was edited and used again in the 2015 spring data collection.

The survey included 3 parts:

- Questions about the Teacher: This section included questions pertaining to demographics, experience, licensure, and education.

- Questions about the Classes: This section asked teachers about all of the math classes in which they had at least 1 participating student. For each class period, teachers were asked about demographic makeup, ability grouping by class period, skill level of the class, and textbook use.

- Questions about the Students: For each participating student, teachers answered questions about students’ individual math instruction received, ability grouping within class, math skill and interest level, math-specific difficulties, and self-regulatory skills within the classroom.

**Parent Interview**

During the project’s second year, parents were asked to participate in a 12-item phone interview. The interview, adapted from the National Household Education Survey (2014), asks questions about household (e.g., how many adults live in the home, mother’s education level), homework (e.g., how often the child receives help with homework),
parent relationships (e.g., how well informed parent is about what child is being taught in math), and parent perceptions (e.g., how well parent did in math, how important parent believes math is for child’s future). Most interviews were conducted with parents in English. However, a small subset of participants only spoke Spanish; interviews with these parents were conducted in Spanish by a bilingual staff member.

Early Scale-Up Project Measures

Direct Assessments

Students in the original pre-k study were directly assessed with a battery of achievement measures. Assessments were conducted by trained members of the research staff and were done in two individual sessions lasting around 30 minutes each. Students were assessed twice during pre-k and at the end of kindergarten and 1st grade.

- Elementary Math Assessment (EMA). The EMA (Clements, Sarama, & Liu, 2008) is a proximal measure of children’s early math skills, one that is closely aligned with the Building Blocks curriculum and designed by the curriculum developers. It includes both number sense and geometry/measurement components which were combined through Rasch modeling to yield one total score (Clements et al., 2008). The EMA was originally developed for prekindergarten but items were added to extend its suitability to kindergarten and first grade. The EMA assesses a child’s developmental progression in skills like verbal counting, subitizing, number comparison, number composition, shape recognition, patterning, spatial imagery, geometric measurement, etc.

- Woodcock Johnson III Tests of Achievement (WJ). The WJ (Woodcock, McGrew, & Mather, 2001) is a standard assessment of a range of skills, designed to be used with people ages 2 to 90+. Three subtests of the WJ were used at all time points in this scale-up study, including two measures of early mathematics skills and one measure of early literacy skills.
  - The Applied Problems subtest assesses children’s ability to solve numerical and spatial problems presented verbally with accompanying pictures of objects.
  - The Quantitative Concepts subtest is described above as it is also being used in the follow-up study.
  - The Letter-Word Identification subtest assesses children’s letter and word identification ability. This literacy subtest was included in the assessment battery primarily to ensure that teachers in the experimental condition who were focusing much more on math instruction were not sacrificing literacy instruction or doing harm to their students’ early reading skill development for the sake of a new math focus.
    - In kindergarten and first grade, the Story Recall subtest was added to the WJ battery. Story Recall assesses children’s narrative recall by requiring them to answer questions about stories the assessor reads aloud.
Renfrew Bus Story - North American Edition (BUS). The BUS (Glasgow & Cowley, 1994) is a tool used to assess young children's narrative retelling skills. It was administered twice during the pre-k year, but not in subsequent years. The BUS yields scores pertaining to children's language abilities as well as information retained from the story.

Teacher Ratings
At the beginning and end of the pre-k year and the end of each of the two subsequent early study years, teachers were asked to rate each child using three different instruments.

- Academic and Classroom Behavior Record (ACBR). The ACBR (Farran, Bilbrey, & Lipsey, 2003) was used to assess the teacher's perception of child learning in various subject matter areas and the child's attitude toward school. Five subscale scores derived from the ACBR are Prosocial Behavior, Withdrawn Behavior, Aggressive Behavior, Preparedness for the Current Grade (except for at the end of pre-k, where the subscale was Preparedness for Kindergarten), and Peer Relations. The Withdrawn Behavior and Aggressive Behavior subscales were reverse-scored in such a way that makes them comparable to the other subscales, where higher scores indicate less negative behavior (i.e., a better score).
- Instrumental Competence Scale for Young Children-Short Form (ICS). The ICS (Lange & Adler, 1997) includes 18 items asking teachers to rate children's self-regulation, motivation, and social assertiveness as displayed in the classroom.
- Cooper-Farran Behavioral Rating Scale (CFBRS). The CFBRS (Cooper & Farran, 1991) items include both social skills and work-related skills. Only the work-related questions were used in this study, which assess children's attentiveness, ability to follow directions, and task persistence.

Parent Questionnaire
In the fall and spring of the pre-k year, we asked parents of each participating child to complete questionnaires (TRIAD, 2003).

- Fall Questionnaire: This first questionnaire contained items grouped into:
  - Skills and Home Environment: Questions involved the child’s math skills, computer games played at home, TV watched (type and amount), and children's books in the home.
  - Family Activities: Parents were asked how often family members interact with the child of interest for a variety of activities including telling stories, singing songs, reading books, counting objects, doing math activities, playing with electronic toys, etc.
  - Demographics: We asked parents how many children lived in the household, what their highest level of education was, what type of work they did, what their average income was, how much their child weighed at birth, etc.

- Spring Questionnaire: The second questionnaire was much briefer and only asked questions about the child's school experiences. We asked how often the teacher
sent home math activities, how prepared those activities made the parent feel to teach their child math, how well the child was doing in math at school, and how late/absent the child was in the last month.
References


The Treatment and Research Institute for Autism Spectrum Disorders (TRIAD). (2003). *Parent questionnaire*. Vanderbilt University, Nashville, TN.
