CEME Technical Report

CEMETR-2018-01
MARCH 2018

The Center for Educational Measurement and Evaluation

Examining the Psychometric Properties of the North Carolina Kindergarten Entry Assessment

Richard G. Lambert

CEME
The Center for Educational Measurement and Evaluation
College of Education
UNC CHARLOTTE

Examining the Psychometric Properties of the North Carolina Kindergarten Entry Assessment

Richard G. Lambert, Ph.D.

Center for Educational Measurement and Evaluation

UNC Charlotte

March, 2018

Executive Summary

This report contains reliability and validity evidence for the North Carolina Kindergarten Entry Assessment system (KEA) using a statewide sample of kindergarten children. The KEA was designed as a formative, developmental, authentic, and criterion referenced classroom resource for teachers. As a formative assessment, the KEA focuses on the learning process and is used to support learning while learning is taking place. It is a tool to support teachers and students, and helps provides valuable feedback to inform and adjust both teaching and learning.

The information that the KEA progressions provide was evaluated as used by North Carolina kindergarten teachers during the fall of 2016 and 2017. The information from each progression was combined into an underlying composite score for psychometric research purposes only. This report provides feedback on the functioning of the KEA with an eye toward improving the use of the measure throughout the state. The data for this study came from all elementary schools within North Carolina that were participating in the KEA assessment system during the 2016-17 and 2017-18 academic years.

The 2016 data came from 115 school districts, 1,251 school sites, over 5,600 teachers, and 106,337 children. The 2017 data came from 115 school districts, 1,237 school sites, 6,439 teachers, and 102,879 children. In 2016, Kindergarten teachers assessed the children in their classrooms using six required progressions: Engagement in Self-selected Activities, Emotional Literacy, Grip and Manipulation, Object Counting, Letter Naming, and Following Directions. Teachers also used at least one of two additional progressions: Hand Dominance and Crossing the Midline. All eight progressions were included in the analyses. In 2017, all eight progressions were required.

Rasch scaling, the one parameter Item Response Theory model, was used to create overall ability estimates for each child and to examine the measurement properties of the information provided by each progression. Data were analyzed using the Partial Credit Model (PCM; Masters, 1982), with Winsteps software (Linacre, 2012). For each progression, the category labels were converted to numeric values for the purpose of the analyses. The "Emerging" category was assigned a value of 0, A was assigned a value of 1, B became 2, C became 3, etc. The highest category, "Beyond", representing abilities beyond the highest behavioral anchors, was assigned a numeric value of 1 greater than the highest lettered category.

Principal Components Analysis of Residuals showed that the total score explained the majority of the variance in the data (2016 - 68.6%, 2017 – 67.1%) in the placements on the progressions. No contrasts accounted for a substantial amount of variance. However, when the first contrast was examined further, there was some evidence that Letter Naming and Object Counting might comprise a possible second factor, with the progressions focused on physical and social development comprising the first factor. This evidence was weak and needs to be monitored in the future as more progressions are implemented.

The fit statistics for all of the progressions were well within acceptable limits. The infit mean square values ranged from 0.89 to 1.19 in 2016 and .88 to 1.27 in 2017. The outfit mean square values ranged from 0.89 to 1.25 in 2016 and .88 to 1.32 in 2017. The progression to total score correlations, with each progression excluded from the total score, were all moderately high (2016 - .55 to .79, 2017 - .57 to .78). These model fit statistics when taken together generally suggest that the data does

fit the Rasch PCM very well. These results also indicated that the data satisfied the assumption that the progressions measure one underlying dimension (global development of the whole child).

The item difficulty indexes for each progression were calculated using classical and modern measurement methods. There were no substantial differences between male and female children for item difficulty. There were also no substantial differences between African American, Hispanic, or white children for item difficulty. The only exception was that Hispanic children scored slightly lower on Letter Naming. To examine for possible differential item functioning, the Rasch item difficulty levels were compared between male and female children, and between African American, Hispanic, and white children. There were no substantial differences in item difficulty levels across these subgroups, by gender or race / ethnicity, for any of the progressions.

The Rasch difficulty of each progression was estimated in logit units. The progressions pertaining to a child's ability to cross the midline and demonstrate hand dominance were estimated as the relatively easiest progressions (2016 = -.52, -.51, 2017 = -.45, -.56). The progressions pertaining to a child's ability to follow directions, engage in self-selected activities, name letters, and demonstrate grip and manipulation skills were found to be of average difficulty level (2016 = -.27 to .17, 2017 = -.21 to .18). The progressions pertaining to object counting and a child's ability to demonstrate emotional literacy were to be the most difficult (2016 = .50, .57, 2017 = .41, .54). Therefore, the developmental pathway that is formed indicates a pathway from the easiest to the most difficult progressions that generally aligns with expectations from developmental theory.

The range of progression difficulty estimates was much narrower than the range of child ability estimates. However, that the range of progression difficulties was effectively much wider than the results indicate when considering the separation created between children by the range of rating scale anchor point threshold locations. Andrich thresholds were estimated using the Rasch PCM. These values indicated the ability locations that form the model estimated boundaries between the rating scale or progression categories. These locations indicate where on the total score the probability becomes higher that a child will be placed at the next highest category on the progression, relative to the previous anchor point. The values were as follows: Engagement (2016 = -3.08 to 3.76, 2017 = -3.00 to 3.76), Object Counting (2016 = -2.55 to 3.50, 2017 = -2.97 to 3.56), Emotional Literacy (2016 = -3.32 to 4.00, 2017 = -3.30 to 3.98), Grip and Manipulation (2016 = -2.98 to 3.81, 2017 = -3.11 to 3.85), Crossing the Midline (2016 = -2.38 to 3.94, 2017 = -2.52 to 4.01), Following Directions (2016 = -3.19 – 3.78, 2017 = -3.22 to 3.89), Letter Naming (2016 = -1.96 to 3.14, 2017 = -1.99 to 3.36), and Hand Dominance (2016 = -2.79 to 3.85, 2017 = -3.09 to 3.80). These values much more closely match the full range of ability estimates on the total score and provide reasonable separation of children according to underlying global development of the whole child.

The item (progression) reliability values for both years, both sample-based and model-based, were greater than .99. The item (progression) separation indexes were also very high (sample-based, 2016 = 99.10, 2017 = 110.9 and model-based, 2016 = 99.57, 2017 = 111.6). Taken together, these findings indicate it is reasonable to expect highly consistent estimates of progression difficulty levels across samples. The sample-based person separation index was 2.64 in 2016 and 2.76 in 2017. The model-based value was 3.05 in 2016 and 3.16 in 2017. The sample-based person reliability index was .87 in 2016 and .88 in 2017, and the model-based value was .90 in 2016 and .91 in 2017. The Cronbach's alpha value for the total score was .84 in 2016 and .86 in 2017. Based on these reliability indexes, the total scores appear to yield adequately reliable information from this sample. It is

important to note that these results address reliability issues related to the use of a total score only and may be very different from the results of an inter-rater reliability study.

Not all rating scales are created equal and not all raters use rating scales effectively. Ideally, rating scale data is most valid when the intended meaning of each of the individual rating scale anchor points is communicated clearly and unambiguously to raters (in this case teachers), and raters use the scales as intended. In the case of the KEA, valid placements can only occur when teachers understand the purpose of formative assessment, are well trained, understand the intended content of both the progressions and their rating scale anchor points, and collect and analyze valid evidences to support placements on the progressions. An examination of rating scale effectiveness can help identify potential problems with the progressions or their use. This study focused on the following research questions in an effort to begin to understand the rating scale category effectiveness of the KEA progressions as used by North Carolina kindergarten teachers:

1.) What are the characteristics of the distributions of placements on each of the progressions?
2.) Do the mean total scores of the children placed in each category increase monotonically along the rating scale for each of the progressions?
3.) Do the thresholds between rating scale categories increase monotonically along the rating scale for each of the progressions?
4.) Do the category probability plots indicate distinct probability distributions for each rating scale point for each of the progressions?

To address research question one, the center, shape, and spread of the distribution of placements for each progression was examined for both years of data with very similar results. Each distribution was reasonably unimodal and symmetrical with several notable exceptions. For Engagement, there were relatively few placements in the extreme (lowest or "Emerging" and highest or "Beyond") categories. For Object Counting, the extreme categories very also used relatively infrequently as was category F. For Emotional Literacy, the lowest category was used relatively infrequently as was category F. For Grip and Manipulation, the distribution was negatively skewed, and the lowest category was relatively infrequently used. For Crossing the Midline, the distribution was also negatively skewed, and the lowest category was relatively infrequently used. For Following Directions, both the lowest category and category F were relatively infrequently used. For Letter Naming, the distribution was negatively skewed, and the lowest category was relatively infrequently used. For Hand Dominance, both the lowest category and category C were relatively infrequently used.

One indicator of the validity of a formative assessment is the extent to which the average total scores of children placed at each successive rating scale category on the progressions increase across the rating scale categories. This issue was examined to address research question two. The observed average total scores did increase as expected across all rating scale categories for each of the progressions for both years of data.

To address research question 3, the Andrich category thresholds were examined for both years of data. For three of the progressions, Engagement, Grip and Manipulation, and Crossing the Midline, all thresholds increased monotonically as expected. However, the remaining five progressions all had at least one disordered threshold. For Object Counting, categories D and G had disordered thresholds. For Emotional Literacy, categories G and I had disordered thresholds. For Following Directions, categories D, E, and J had disordered thresholds. For Letter Naming, categories C, D, F, and H had disordered thresholds. For Hand Dominance, categories B and D had disordered

thresholds. Each of these identified categories should be examined further as disordered thresholds can present a threat to the validity of the rating scale data and any inferences made from the data.

To address research question 4, category probability plots were examined. These plots indicate the probability distribution for a child being placed on a particular response category, or level on each developmental progression, given their overall ability. The plots should contain distinct and minimally overlapping probability distributions for each rating scale category. Overall, these plots suggested potential difficulties with the use of the rating scales. All eight progressions had at least one category with substantial overlap with adjacent categories. These results may suggest a need to improve the definitions of the category anchors, improve teacher training and understanding of the distinct differences between categories, a need for teachers to collect higher quality evidences, or some combination of these factors.

These analyses focused on only two years of KEA implementation and for many teachers and schools, the process of getting to full implementation is still ongoing. The 2016-17 academic year was the second year of full state implementation of the KEA and 2017-18 was the third. Many teachers and administrators are still becoming familiar with the KEA progressions and assessment process. However, the results of the analyses related to dimensionality and reliability are all very strong and reflect very positively on the use of a total score for psychometric research purposes such as those outlined in this report.

In summary, the results related to item or progression difficulty estimates were generally positive, though they suggest a need for a greater range of progression difficulties. There were no indications of differential item functioning by subgroups based on gender or race / ethnicity. The distributions of scores from all of the progressions were moderately correlated with each other. As expected, the lowest correlations were between progressions that would not be expected to be highly related (i.e. Letter Naming and Hand Dominance, 2016 $r = .341$, 2017 $r = .393$). Similarly, the highest correlations were between progressions that would be expected to be related (i.e. Emotional Literacy and Following Directions, 2016 $r = .610$, 2017 $r = .634$).

The results of the examination of rating scale category effectiveness were mixed. There are some very positive results, such as those related to the expected increases in total scores across the categories. There are some mixed results, such as those related to distributional shape and use of the complete scales. There are also some results that indicate cause for concern related to the need for more distinct category probability plots. Future research and continued examination of the psychometric properties of the developmental progressions will be needed to monitor ongoing progress toward full implementation of the KEA assessment as it was intended to be used, and to determine the sources of the continuing challenges for teachers.

Examining the Psychometric Properties of the North Carolina Kindergarten Entry Assessment

This report focuses on establishing reliability and validity evidence for the North Carolina Kindergarten Entry Assessment system (KEA). The KEA yields information that is rooted in the ongoing every day work of teachers. Teachers collect ongoing portfolios of evidences, reflect upon and analyze those evidences, make preliminary ratings on an ongoing basis, and finalize ratings on a series of developmental progressions at the end of a 60 day assessment period at the beginning of the kindergarten year. This information is intended to be used to inform instruction and to facilitate communication with parents and other stakeholders. In contrast to direct assessments, evidences are collected within regular activities in natural classroom contexts and help teachers understand and observe child progress, plan instruction, and scaffold and support child growth and development. In addition, the process of evidence formation and collection directly involves young children in dialogue with teachers about their developmental progress.

The measurement properties of any assessment system should be rigorously examined as long as the measure is in use and the results made available to stakeholders. This process needs to extend to any and all subgroups of children and specific uses of the measure. Reliability and validity are not inherent qualities of an assessment, but rather are properties of the information an assessment provides under particular conditions of use. It is particularly important to provide teachers of young children formative assessment measures that are reliable, valid, and culturally sensitive. This report examines and extends the reliability and validity of the assessment evidence for the KEA using a statewide sample of kindergarten children.

The Purpose of the NC KEA

The KEA has been designed and validated to be used as a formative, developmental, authentic, and criterion referenced classroom measurement tool for teachers. By extension therefore,

it is not a screener, summative, benchmark, direct, or norm referenced assessment tool. The primary purpose of the assessment system is to provide teachers with instructionally relevant information about the children they teach. As with any assessment tool, users must always keep in mind the central purpose of a measure, and select appropriate processes that match the purpose of any assessment task. Therefore, it is valuable for teachers and administrators to become aware of the appropriate and inappropriate uses of the KEA and the information it provides.

First and foremost, KEA is a formative assessment. Formative assessment focuses on the learning process and is used to support learning while learning is taking place. Formative assessment has been defined as "…a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes…" (Linquanti, 2014; AERA/APA/NCME, 2014). The KEA can be a very helpful resource when teachers use it to get to know children at the beginning of the school year. It can help teachers understand the strengths that each child brings to the classroom and the specific areas where each child needs support.

The KEA consists of a series of developmental progressions. When teachers communicate with parents, formative assessment data can help them do so in terms that can be easily accessed and understood. Teachers can point parents to placements on the developmental progressions and associated child work samples and anecdotes that address child progress with specific examples of what children know and can do. Formative assessment information is also particularly helpful for teachers when they communicate and collaborate with other educational professionals within their professional learning communities. Data and evidence driven conversations can lead to richer interactions with everyone connected to the children. A rich and detailed picture of a child's current learning status and their patterns of growth and development can help other educational professionals provide individualized and informed support to the child. Teachers can use these

richer conversations to solicit the participation of involved professionals in the evidence gathering process, and can gather additional understanding of each child as they seek specific input from educational professionals about how to support children.

As useful as formative assessment information and processes can be to teachers, formative assessment is not summative assessment. It is not appropriate to use the information provided by formative assessments about specific children or groups of children for any summative purposes such as performance evaluation of teachers, program evaluation, or assessment of classroom, center, or program quality. It is also inappropriate to use the information yielded by formative assessments to make any kind of high stakes decisions. In fact, attempting to do so can give teachers perverse incentives to make less than valid placements on the developmental progressions and can thereby rob them and the children they serve of the benefits of the appropriate uses of formative assessment information.

When teachers have a more complete understanding of a child's developmental pathway toward accomplishing specific instructional objectives, they can comprehend more clearly what is the next step for each child. They can then use that enhanced understanding to plan instruction, enrich communication with parents and stakeholders, and inform everyday interactions with the child. Perhaps most importantly, they can use this understanding to help provide meaningful feedback to children, helping them understand what skills and abilities will be emerging next. This process can help children become more engaged in and excited about learning, and can give them a more meaningful sense of accomplishment during the learning process. This process can also help children become intentional participants in the assessment of their own learning and development, contributing evidences of their choosing to their merging portfolios. Children can then become more involved in the self-regulation of their own learning and self-assessment, and can more fully receive, understand, and utilize teacher and parent feedback about their progress.

The KEA has been designed to be an authentic assessment. Authentic assessment resources help teachers observe the progress children are making through a process of gathering evidences of learning that emerge naturally from within daily classroom activities. These evidences are intended to be gathered within regularly occurring instructional activities and routines. The information that the KEA provides is intended to be rooted in these ongoing processes through which teachers gather rich portfolios of evidences of student growth, analyze those evidences, make periodic placements on developmental progressions based on those evidences, and use those placements to plan and support the next steps in the learning process. In this way, the KEA supports assessment "for" learning and assessment "about" the learning process, and not just assessment "of" the results of learning (Heritage, 2013).

Authentic assessment is not direct assessment. Direct assessments include standardized protocols of assessment activities that "done to" a child. This means that children are presented with specific assessment prompts or question formats that are designed to elicit specific correct or incorrect responses from children. Direct assessment takes place in an intentionally created artificial testing situation, rather than in the course of daily activities. Direct assessments are appropriate measures for some testing purposes and are widely and correctly used within the broader educational system, particularly with children older than the early childhood years. They can play important roles within a comprehensive assessment system and are appropriately used when objective, summative, data are required concerning how individual children or groups of children are functioning at a particular point in time.

The authentic process used for formative assessments has often been described as a continuous cycle of activities that is part of everyday instructional activity in the classroom. This cycle is often outlined in phases: 1.) understanding what is next for a child and set learning goals, 2.) defining and understanding criteria that will indicate progress toward the next level of development,

3.) gathering evidences of growth, development and learning, 4.) analysis and interpretation of evidences, 5.) making placements on developmental progressions, and 6.) adapting instruction to support the unique needs of the individual child (Heritage, 2013). This cycle can then repeat itself as the child moves toward the next developmental level on a specific progression related to an instructional objective. This process is also simultaneously playing out over many developmental progressions across a variety of learning objectives and developmental domains. This cycle begins with a data-driven sense of where a child is currently functioning relative to a particular developmental pathway, and progresses through to data-driven support for the growth, learning, and development of the child. It is an integral part of the instructional process and is neither distinct from nor supplemental to learning. Rather, it is the natural manifestation of high quality instructional practices and enhances the teacher's understanding of a child's current developmental status, progress over time, and needs for support. It also provides systematic steps through which teachers can strengthen their feedback to children and communication with parents and other educational professionals.

Finally, the KEA, like all formative assessment measures, has been designed to facilitate a dynamic process that is fully integrated into and at the center of the teaching and learning process (Shepard, 2000). It has been designed to directly support the learning and development of children. Therefore, the information that the individual progressions provide will be most useful to teachers if it stands alone and can be directly translated into enhanced teacher understanding of children and their instructional needs. This report is designed to evaluate the information that the progressions provide as used by North Carolina kindergarten teachers. The information from each progression has been combined into an underlying composite score for psychometric diagnostic and research purposes only and this report is produced to provide feedback on the functioning of the KEA as it

was used in practice during the 2016-17 and 2017-18 academic years with an eye toward improving the use of the measure throughout the state.

The Data Source

The data for this study came from all elementary schools within North Carolina that were participating in the KEA assessment system during the 2016-17 and 2017-18 academic years. In 2016, Kindergarten teachers were required to assess the children in their classrooms using six required progressions and at least one of two additional progressions. In 2017, all eight progressions were required. All eight progressions were included in the analyses for both years. The 2016 data set included 115 school districts, 1,251 school sites, and 106,337 children. It was not possible to obtain the exact number of teachers participating as Wake County stored their data in a separate format from the other districts around North Carolina and did not include teacher identifiers. However, our best estimate is that approximately 5,600 teachers participated. The exact demographic characteristics of the population of teachers and children used in this study could not be determined as not all demographic variables were available. However, the data set is so large and comprehensive that there is no reason to believe that those characteristics would differ from those of the actual population of North Carolina kindergarten children and their teachers. The 2017 data came from 115 school districts, 1,237 school sites, 6,439 teachers, and 102,879 children.

Analyses Related to the Construction of the Total Scale Score

For both the 2016 and 2017 data, Rasch scaling, the one parameter IRT model, was used to create overall ability estimates for each child and to examine the measurement properties of the information provided by each progression. Data were analyzed using the Partial Credit Model (PCM; Masters, 1982), with Winsteps software (Linacre, 2012). The Rating Scale (RSM; Bond & Fox, 2001) and the PCM are the two most widely used Rasch model for polytomous response data. The PCM, rather than the RCM, was chosen because the progressions do not share the same rating scales (i.e.,

use of the same number of rating scale categories and labels across progressions). In cases where each progression has its own rating scale structure, the PCM is the appropriate model to apply. Specifically, the 8 KEA progressions required for both the 2016-17 and 2017-18 academic years included in the analyses were the following: Engagement in Self-selected Activities, Object Counting, Emotional Literacy, Grip and Manipulation, Crossing the Midline, Following Directions, Letter Naming, and Hand Dominance. For each progression, the category labels were converted to numeric values for the purpose of the analyses. The "Emerging" category was assigned a value of 0, A was assigned a value of 1, B became 2, C became 3, etc. The highest category, "Beyond", representing abilities beyond the highest behavioral anchors, was assigned a numeric value of 1 greater than the highest lettered category. Therefore, for the purposes of the analyses, the progressions had the following scaling: 1 progression included a 0-4 scale, 2 progressions included a 0-5 scale, 1 progression included a 0-6 scale, 1 progression included a 0-9 scale, 1 progression included a 0-10 scale, and two progressions included a 0-12 scale.

Results from the fall, 2016 Assessment

Dimensionality

Rasch modeling assumes what is called unidimensionality, meaning that the progressions in question measure one and only one underlying latent construct. In the case of the KEA, this latent construct might be considered global development of the whole child. The unidimensionality of the total score, or scale, was evaluated by using Mean Square (MNSQ) progression fit statistics and Rasch Principal Components Analysis of residuals (PCAR). The MNSQ fit values between 0.6 and 1.4 are considered reasonable for rating scale progressions (Bond & Fox, 2007). MNSQ values less than 2.0 can indicate that a progression, though not fitting optimally with the measurement model, can still contribute useful information to the overall score on the measure. Progressions with mean

square values of between 1.4 and 2.0 can be considered potentially unproductive for the construction of measurement scales, but not degrading to the quality of the information provided by the scale (Linacre, 2002). Infit statistics indicate the fit of individual progression response patterns to the measurement model. They also address the possibility of secondary dimensions and fit to the underlying construct. Outfit statistics are sensitive to outliers; that is responses that show great differences between person responses and progression difficulties. They are also sensitive to unusual and unexpected progression response patterns.

For PCAR, a variance of greater than 50% explained by measures is considered good, and offers support for scale unidimensionality. If a secondary dimension has an eigenvalue of smaller than 3 and accounts for less than approximately 5% of the unexplained variance, unidimensionality is considered plausible (Linacre, 2012).

The PCAR showed that the Rasch dimension explained the majority of the variance in the data (68.6%) with an eigenvalue of 17.5, relative to the total eigenvalue of 25.5. The first contrast (the largest potential secondary dimension) had an eigenvalue of 1.6 and accounted for 6.3% of the unexplained variance. When the first contrast was examined further, there was some evidence that Letter Naming and Object Counting might comprise a possible second factor, with the progressions focused on physical and social development comprising the first factor. However, this evidence was weak and needs to be monitored in the future as more progressions are implemented.

The fit statistics for all of the progressions were well within acceptable limits (see Table 1). The infit MNSQ values ranged from 0.89 to 1.19. The outfit MNSQ values ranged from 0.89 to 1.25. The progression to total score correlations, with each progression excluded from the total score, ranged from .55 to .79. The progression to total score correlations, with each progression included in the total score, ranged from .57 to .79. In summary, these model fit statistics when taken

together generally suggest that the data does in fact fit the Rasch PCM very well. These results also indicated that the data satisfied the unidimensionality assumption of the Rasch model.

Item Difficulty Measures

The progression location hierarchy appeared to be generally consistent with the expected developmental trajectory for typically developing kindergarten children. Table 1 lists the progression difficulty estimates from highest to lowest along with the standard errors for these estimates and the associated fit statistics. These results were evaluated using the final data available at the end of the fall 60-day KEA assessment time period. The progressions pertaining to a child's ability to cross the midline and demonstrate hand dominance were estimated as the relatively easiest progressions (-.52, -.51). The progressions pertaining to a child's ability to follow directions, engage in self-selected activities, name letters, and demonstrate grip and manipulation skills were found to be of average difficulty level (-.27 to .17). The progressions pertaining to object counting and a child's ability to demonstrate emotional literacy were to be the most difficult (.50, .57).

The range of progression difficulties (-.52 to .57) was found to be relatively narrow and it will be ideal to add progressions with a wider range of difficulty levels in the future. This can be seen in Figure 1. This figure displays the Item Person Map. On the left side of the center of the map, the distribution of total scores for the population of children is displayed. This distribution conforms closely to a unimodal and symmetrical shape and indicates that the total measure score is functioning well to spread children out according to underlying overall developmental status. The right side of the map indicates the location of each progression. The progression locations, or difficulty estimates, indicate that the progressions are functioning well to separate children near the center of the distribution and are less useful for spreading out children at the upper and extreme lower ends of the distribution. The practical implication for teachers is that these eight progressions may be

relatively less useful for understanding and supporting the developmental progress of children with very low or more highly developed global development across domains.

When the progression rating scale anchor point, or category, locations are considered, these values come closer to matching the range of abilities of the children assessed. In tables 2 through 9, the Andrich thresholds are reported. These values indicate the ability locations that form the model estimated boundaries between the rating scale or progression categories. These locations indicate where on the underlying ability scale, or total score, the probability becomes higher that a child will be placed at the next highest category on the progression, relative to the previous anchor point. The values were as follows: Engagement = -3.08 – 3.76, Object Counting = -2.55 – 3.50, Emotional Literacy = -3.32 – 4.00, Grip and Manipulation = -2.98 – 3.81, Crossing the Midline = -2.38 – 3.94, Following Directions = -3.19 – 3.78, Letter Naming = -1.96 – 3.14, and Hand Dominance = -2.79 – 3.85. These values more closely match the full range of ability estimates on the total score and provide reasonable separation of children according to underlying ability.

In summary, the developmental pathway that is formed indicates a pathway from the easiest to the most difficult progressions that generally aligns with expectations from developmental theory. It is also important to recognize, as indicated, that the range of progression difficulties is effectively much wider than the results indicate when considering the separation created between children by the range of rating scale anchor point threshold locations.

Reliability

Reliability was evaluated using the following Rasch indexes: the person separation index, item separation index, person reliability, and item reliability. Item (progression) and person reliabilities were evaluated using both sample-based and model-based coefficients. The person separation index, an estimate of the adjusted person standard deviation divided by the average measurement error, indicates how well the instrument can discriminate persons on each of the

constructs. The item (progression) separation index indicates an estimate in standard error units of the spread or separation of progressions along the measurement constructs. Reliability separation indexes greater than 2 are considered adequate, and indexes greater than 3 are considered high (Bond & Fox, 2007). High person or item (progression) reliability means that there is a high probability of replicating the same separation of persons or progressions across measurements. Specifically, person separation reliability estimates the replicability of person placement across other progressions measuring the same construct. Similarly, progression separation reliability estimates the replicability of progression placement along the construct developmental pathway if the same progressions were given to another sample with similar ability levels. The person reliability provided is similar to the classical or traditional test reliability whereas the progression reliability has no classical equivalent. Low values in person and progression reliability may indicate a narrow range of person or progression measures. It may also indicate that the number of progressions or the sample size under study is too small for stable estimates (Linacre, 2009). Reliability was also evaluated using Cronbach's alpha measure of internal consistency.

The item (progression) reliability values, both sample-based and model-based, were greater than .99. The item (progression) separation indexes were also very high: sample-based = 99.10 and model-based = 99.57. Taken together, these findings indicate it is reasonable to expect highly consistent estimates of progression difficulty levels across samples. The sample-based person separation index was 2.64 and the model-based value was 3.05. The sample-based person reliability index was .87 and the model-based value was .90. The Cronbach's alpha value for the total score was .84. Based on these reliability indexes, the total scores appear to yield adequately reliable information from this sample. Specifically, these results indicate that it is reasonable to expect reliable estimates of child overall ability levels when teachers use the KEA to place kindergarten children along the developmental progressions, and those individual progression scores are transformed into a

composite or total score. It is important to note that these results address reliability issues related to the use of a total score and may be very different from the results of an inter-rater reliability study.

Rating Scale Category Effectiveness

A rating scale with demonstrated category effectiveness yields evidence that raters are using the scale as it was intended to be used. This means that raters can use the scale to discriminate between responses with true underlying differences on the construct being measured. In the case of the KEA, rating scale category effectiveness is a measure of the validity of the data elicited by the developmental progressions. Developmental progressions with effective rating scales yield valid data that can be used to place children along a continuum of development so that the placements both reflect the true developmental status of each child and can be used by teachers to differentiate instruction and support growth, learning, and development. Therefore, the rating scale category effectiveness of the KEA was examined to provide information about the rating scale categories on specific progressions and to evaluate whether teachers appear to be using the progressions in the manner intended. Rating scale effectiveness was also examined to evaluate if it is reasonable to apply Rasch modeling to the data.

Not all rating scales are created equal and not all raters use rating scales effectively. Ideally, rating scale data is most valid when the intended meaning of each of the individual rating scale anchor points is communicated clearly and unambiguously to respondents or raters, and raters use the scales as intended. The evaluation of rating scale category effectiveness can suggest the optimal number of rating categories, places along the scale where categories can be combined, and categories that may be misunderstood or misused by raters. In the case of the KEA, valid placements can only occur when teachers understand the purpose of formative assessment, are well trained, understand the intended content of both the progressions and their rating scale anchor points, and collect and analyze valid evidences to support placements on the progressions. An examination of rating scale

effectiveness can help identify potential problems with the progressions or their use. However, further research is often needed to determine whether identified problems are related to the progressions themselves, their use by raters, the quality of rater training, or some combination of these factors.

This study focused on the following research questions in an effort to begin to understand the rating scale category effectiveness of the KEA progressions as used by North Carolina kindergarten teachers:

5.) What are the characteristics of the distributions of placements on each of the progressions?
6.) Do the mean total scores of the children placed in each category increase monotonically along the rating scale for each of the progressions?
7.) Do the thresholds between rating scale categories increase monotonically along the rating scale for each of the progressions?
8.) Do the category probability plots indicate distinct probability distributions for each rating scale point for each of the progressions?

To address research question 1, the center, shape, and spread of the distribution of ratings was examined for each progression. It is recommended that for each progression, each rating scale category needs to be assigned to a minimum of 10 children. All rating scale categories should be used by the raters and each category should be assigned to enough children to allow for reasonable statistical estimates within the Rasch modeling process. These criteria were easily met for all eight progressions. Across the eight progressions, the full range of categories, from "Emerging" to "Beyond", was used by the teachers. Tables 2 through 9 include the number and percent of children assigned to each rating scale category. Table 10 includes the mean, median, and standard deviation for each progression. The median is also reported as the median lettered category for each progression.

Figures 2 through 9 display the shape of the distribution of ratings for each progression through a simple bar chart of the percentage of children placed in each rating scale category. These charts indicate a reasonably unimodal and symmetrical shape to the distribution of ratings for each progression with several notable exceptions. For Engagement, there were relatively few placements in the extreme (lowest or "Emerging" and highest or "Beyond") categories. For Object Counting, the extreme categories very also used relatively infrequently as was category F. For Emotional Literacy, the lowest category was used relatively infrequently as was category F. For Grip and Manipulation, the distribution was negatively skewed, and the lowest category was relatively infrequently used. For Crossing the Midline, the distribution was also negatively skewed, and the lowest category was relatively infrequently used. For Following Directions, both the lowest category and category F were relatively infrequently used. For Letter Naming, the distribution was negatively skewed, and the lowest category was relatively infrequently used. For Hand Dominance, both the lowest category and category C were relatively infrequently used.

To address research question 2, the average of the overall ability estimates, based on the total progression scores, for all children in the sample who were placed at a particular response category or scale point on each of the developmental progressions was examined. Average measure scores should advance monotonically with rating scale category values (Bond & Fox, 2007). Tables 2 through 9, under the column labeled Observed Average, demonstrate that the average total scores did increase as expected across all rating scale categories for each of the progressions. This finding is a very positive result for the validity of the progressions and is also illustrated graphically in figures 10 through 17.

To address research question 3, the category thresholds were examined. Thresholds (also called step calibrations) are the difficulty levels estimated as the point on the total score at which teachers are more likely to choose one response category or rating scale point over the previous step

on the progression (Bond & Fox, 2007). For this study the Andrich thresholds from the Partial Credit Model were used. Thresholds should also increase monotonically along the rating scale categories. These values are reported in tables 2 through 9 under the column labeled Andrich Threshold. For three of the progressions, Engagement, Grip and Manipulation, and Crossing the Midline, all thresholds increased monotonically as expected. However, the remaining five progressions all had at least one disordered threshold. These are indicated by boxes around italicized threshold values. For Object Counting, category D had a disordered threshold. For Emotional Literacy, categories G and I had disordered thresholds. For Following Directions, categories D, E, and J had disordered thresholds. For Letter Naming, categories C, D, F, and H had disordered thresholds. For Hand Dominance, category B had a disordered threshold. Each of these identified categories should be examined further as disordered thresholds present a threat to the validity of the rating scale data and any inferences made from the data.

To address research question 4, category probability plots were examined. These plots indicate the probability distribution for a child being placed on a particular response category, or level on each developmental progression, given their overall ability or total measure score. The plots should contain distinct and minimally overlapping probability distributions for each rating scale category. The magnitude of the distances between adjacent category thresholds should be large enough so that each step defines a distinct position and each category has a distinct peak in the category probability curve plot (Bond & Fox, 2007). Figures 18 through 25 displays these plots. Overall, these plots suggest substantial difficulties with the use of the rating scales. These may suggest a need to reduce or combine categories, improve the definitions of the category anchors, improve teacher training and understanding of the distinct differences between categories, a need for teachers to collect higher quality evidences, or some combination of these factors. For Engagement, category C shows substantial overlap with adjacent categories. For Object Counting, categories C

through G show substantial overlap with adjacent categories. For Emotional Literacy, categories D through H show substantial overlap with adjacent categories. For Grip and Manipulation, categories C and D show substantial overlap with adjacent categories. For Crossing the Midline the plot indicates appropriately distinct probability distributions. For Following Directions, categories C through J show substantial overlap with adjacent categories. For Letter Naming, categories C through J show substantial overlap with adjacent categories. For Hand Dominance categories A and D show substantial overlap with adjacent categories.

Differences by Subgroup

Another type of evidence for the validity of the information produced by developmental rating scales is extent to which different subgroups of children receive similar scores. Specifically, two children with the same underlying ability should receive the same placement on each developmental progression, and this expectation should be sustained independent of subgroup membership. If children with the same underlying ability receive different placements on the progressions and those differences are systematic based on subgroup membership, then the possibility exists for some level of bias to be inherent in the assessment process. This bias could be related to item content, rater biases, training, or other factors. However, it is unacceptable under any conditions of use. Therefore, subgroup differences based on both gender and race / ethnicity were examined using both classical and modern measurement strategies.

Classical item difficulty was examined by observing the mean score on each progression for the total sample and for the subgroups of interest. Table 12 displays these values. There were no substantial differences between subgroups based on either gender or race / ethnicity. There were only a few exceptions to this finding. White children, on average, tended to be placed somewhat higher (.60 - .77 scale points) than their African American or Hispanic counterparts on the Emotional Literacy and Object Counting progressions. White children, on average, tended to be

placed moderately higher (.81 – 1.21 scale points) than their African American or Hispanic counterparts on the Letter Naming and Following Directions progressions.

These differences in item difficulty were also examined using Rasch modeling. This method investigates the possibility of Differential Item Functioning by examining the item difficulty estimates by subgroup while controlling for underlying ability estimates on the total score. This method, therefore, effectively compares children across the subgroups who have the same underlying total ability estimates. There were no substantial differences between item difficulty estimates based on subgroups using this method. Differences in item difficulty estimates greater than or equal to .64 are considered large, .43 - .63 moderate, and less than .43 are considered negligible (Zwick, Thayer, & Lewis, 1999). The separate item difficulty estimates for males and females are listed in Table 12. The differences between estimates for male and female children, in logit units, ranged from .00 to .18. The separate item difficulty estimates for white, African American, and Hispanic children are also listed in Table 12. The differences between estimates for white and African American children, in logit units, ranged from .00 to .16. The differences between estimates for white and Hispanic children, in logit units, ranged from .00 to .23.

Table 1
Item level statistics and difficulty estimates - 2016

Progression	Item Difficulty	SE	Infit Mnsq	Outfit Mnsq	Item-Measure *r* Item Included	Item Excluded
Emotional Literacy	0.57	< .005	1.01	1.03	0.77	0.77
Object Counting	0.50	< .005	0.96	0.98	0.77	0.76
Following Directions	0.17	< .005	1.00	1.00	0.79	0.79
Engagement	0.05	< .005	0.89	0.89	0.68	0.64
Letter Naming	0.01	< .005	1.19	1.25	0.74	0.76
Grip and Manipulation	-0.27	< .005	0.95	0.94	0.68	0.67
Hand Dominance	-0.51	0.01	1.01	1.01	0.60	0.60
Crossing the Midline	-0.52	0.01	0.97	0.96	0.57	0.55

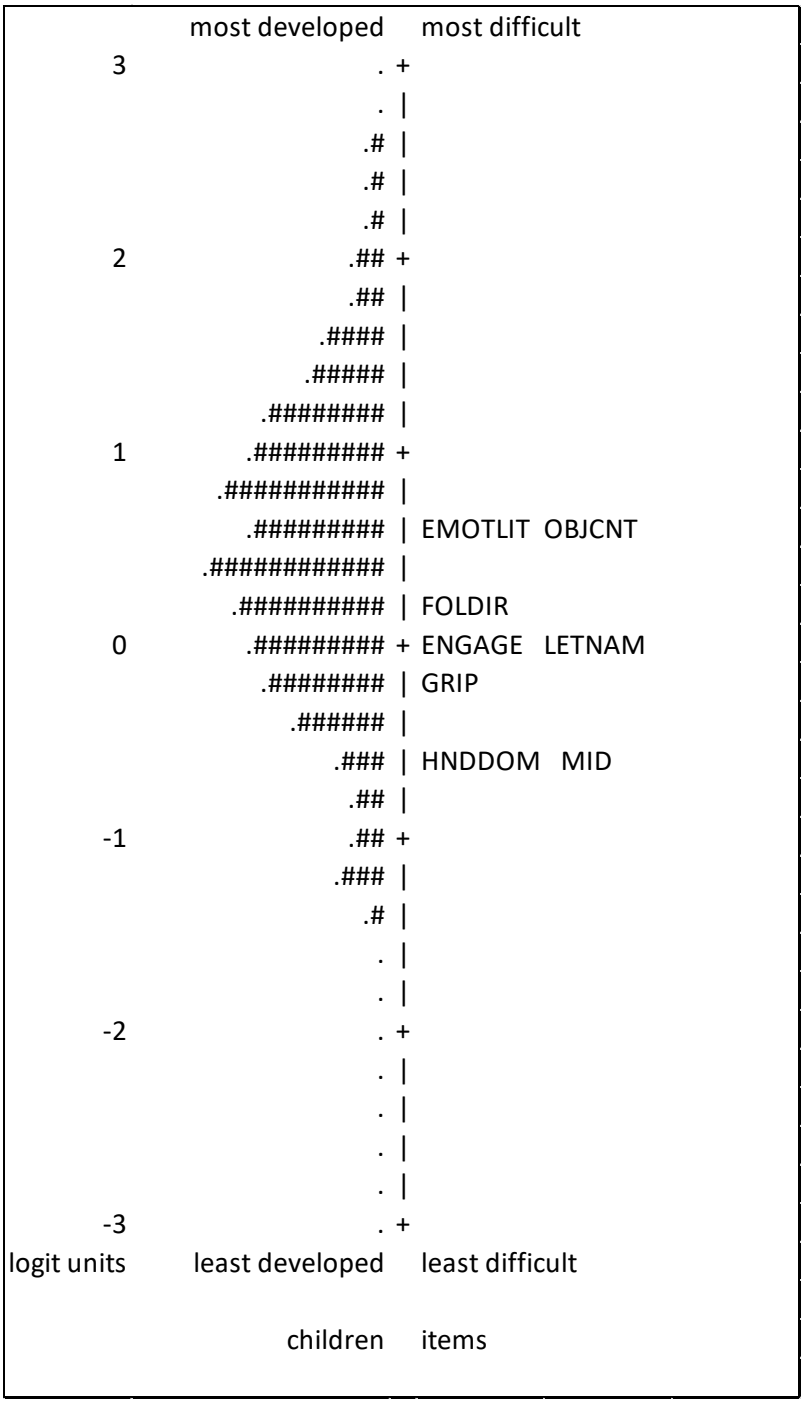

Figure 1. Item Person Map - 2016.

Table 2
Category thresholds and observed average measure scores for Engagement - 2016

Progression Categories	Counts	Percent	Observed Average	Expected	Infit Mnsq	Outfit Mnsq	Andrich Threshold
Emerging	1646	1.61%	-1.85	-1.86	1.04	1.03	-----
A	8974	8.78%	-0.72	-0.69	0.96	0.95	-3.09
B	27752	27.16%	-0.06	0.01	0.85	0.83	-1.49
C	27368	26.79%	0.53	0.53	0.84	0.82	0.25
D	32709	32.02%	1.15	1.09	0.87	0.88	0.57
Beyond	3714	3.64%	2.28	2.15	0.92	0.95	3.76

Table 3
Category thresholds and observed average measure scores for Object Counting - 2016

Progression Categories	Counts	Percent	Observed Average	Expected	Infit Mnsq	Outfit Mnsq	Andrich Threshold
Emerging	2780	2.73%	-1.64	-1.78	1.30	1.18	-----
A	5522	5.43%	-0.97	-0.90	0.90	0.90	-2.55
B	12954	12.74%	-0.42	-0.35	0.87	0.90	-1.96
C	10505	10.33%	-0.01	0.03	0.76	0.77	-0.44
D	22600	22.23%	0.38	0.35	0.90	0.90	***-1.07***
E	14871	14.63%	0.70	0.64	0.90	0.97	0.41
F	8101	7.97%	0.97	0.93	0.93	1.00	0.89
G	10735	10.56%	1.26	1.26	1.03	1.05	***0.31***
H	11437	11.25%	1.69	1.72	1.09	1.10	0.91
Beyond	2147	2.11%	2.72	2.79	1.25	1.14	3.50

Table 4
Category thresholds and observed average measure scores for Emotional Literacy - 2016

Progression Categories	Counts	Percent	Observed Average	Expected	Infit Mnsq	Outfit Mnsq	Andrich Threshold
Emerging	1553	1.53%	-1.89	-2.11	1.53	1.33	-----
A	4359	4.29%	-1.10	-1.16	1.19	1.15	-3.32
B	8020	7.89%	-0.62	-0.53	0.92	0.92	-1.99
C	15642	15.39%	-0.12	-0.08	0.85	0.87	-1.53
D	18130	17.84%	0.27	0.26	0.89	0.90	-0.62
E	21015	20.68%	0.60	0.57	0.94	0.95	-0.29
F	7663	7.54%	0.96	0.87	0.88	0.90	1.16
G	11852	11.66%	1.17	1.18	1.07	1.10	***0.02***
H	5554	5.46%	1.56	1.54	1.04	1.12	1.54
I	6590	6.48%	1.96	2.07	1.27	1.30	***1.03***
Beyond	1260	1.24%	3.45	3.32	1.08	1.08	4.00

Table 5
Category thresholds and observed average measure scores for Grip and Manipulation - 2016

Progression Categories	Counts	Percent	Observed Average	Expected	Infit Mnsq	Outfit Mnsq	Andrich Threshold
Emerging	991	1.00%	-2.22	-2.27	1.20	1.10	-----
A	3967	4.01%	-1.02	-1.13	1.22	1.28	-2.98
B	14648	14.79%	-0.48	-0.39	0.87	0.87	-1.77
C	16769	16.94%	0.11	0.11	0.89	0.85	0.01
D	23342	23.57%	0.52	0.55	0.88	0.83	0.27
E	34632	34.98%	1.10	1.06	0.92	0.93	0.66
Beyond	4664	4.71%	2.02	2.03	1.12	1.03	3.81

Table 6
Category thresholds and observed average measure scores for Crossing the Midline - 2016

Progression Categories	Counts	Percent	Observed Average	Expected	Infit Mnsq	Outfit Mnsq	Andrich Threshold
Emerging	776	1.19%	-1.95	-2.08	1.14	1.27	-----
A	2951	4.53%	-0.71	-0.80	1.06	1.09	-2.38
B	15731	24.17%	-0.09	0.04	0.89	0.86	-1.51
C	41127	63.20%	0.76	0.72	0.89	0.91	-0.06
Beyond	4489	6.90%	1.68	1.70	1.04	1.00	3.94

Table 7
Category thresholds and observed average measure scores for Following Directions - 2016

Progression Categories	Counts	Percent	Observed Average	Expected	Infit Mnsq	Outfit Mnsq	Andrich Threshold
Emerging	975	0.96%	-2.16	-2.48	2.05	1.32	-----
A	2285	2.26%	-1.52	-1.52	1.24	1.08	-3.19
B	6874	6.79%	-1.02	-0.89	0.71	0.77	-2.45
C	6095	6.02%	-0.49	-0.47	0.78	0.79	-0.72
D	10421	10.29%	-0.16	-0.17	0.90	0.87	***-1.02***
E	9415	9.30%	0.12	0.06	0.98	0.93	-0.12
F	5786	5.71%	0.31	0.28	1.01	1.03	***0.49***
G	9624	9.50%	0.48	0.49	1.02	1.05	-0.30
H	17160	16.95%	0.74	0.72	0.99	1.05	-0.15
I	10125	10.00%	1.01	0.99	1.02	1.09	1.21
J	11866	11.72%	1.28	1.31	1.14	1.11	***0.81***
K	8686	8.58%	1.78	1.80	1.19	1.10	1.67
Beyond	1941	1.92%	2.96	2.93	1.25	1.04	3.78

Table 8
Category thresholds and observed average measure scores for Letter Naming - 2016

Progression Categories	Counts	Percent	Observed Average	Expected	Infit Mnsq	Outfit Mnsq	Andrich Threshold
Emerging	3221	3.17%	-1.63	-1.78	1.93	1.49	-----
A	5278	5.20%	-1.09	-1.12	1.46	1.60	-1.96
B	5021	4.94%	-0.85	-0.75	0.87	1.11	-0.88
C	1694	1.67%	-0.46	-0.50	1.01	1.20	***0.46***
D	3376	3.32%	-0.23	-0.31	1.21	1.47	***-1.10***
E	6013	5.92%	-0.08	-0.15	1.20	1.46	-0.82
F	3082	3.03%	0.02	0.00	1.12	1.33	***0.59***
G	4351	4.28%	0.20	0.17	1.12	1.29	-0.27
H	9018	8.88%	0.39	0.35	1.18	1.34	***-0.48***
I	13399	13.19%	0.55	0.58	1.17	1.09	0.05
J	20641	20.32%	0.86	0.87	1.15	1.18	0.27
K	21409	21.08%	1.26	1.28	1.16	1.06	1.01
Beyond	5052	4.97%	2.04	2.15	1.38	1.05	3.14

Table 9
Category thresholds and observed average measure scores for Hand Dominance - 2016

Progression Categories	Counts	Percent	Observed Average	Expected	Infit Mnsq	Outfit Mnsq	Andrich Threshold
Emerging	181	0.42%	-2.13	-2.46	1.20	1.09	-----
A	815	1.89%	-0.94	-0.87	0.97	0.99	-2.79
B	15940	36.89%	-0.02	-0.03	1.00	1.00	***-2.86***
C	3967	9.18%	0.42	0.48	0.85	0.79	2.14
D	19389	44.87%	1.02	1.00	0.98	1.00	***-0.34***
Beyond	2919	6.76%	1.85	1.92	1.18	1.15	3.85

Table 10
Descriptive statistics by progression - 2016

	Engagement	Object Counting	Emotional Literacy	Grip and Manipulation	Crossing Midline	Following Directions	Letter Naming	Hand Dominance
Mean	2.90	4.55	4.82	3.82	2.70	6.83	7.96	3.16
Median	3.00	4.00	5.00	4.00	3.00	7.00	9.00	4.00
SD	1.11	2.23	2.22	1.34	0.71	2.96	3.41	1.08
Median Category	C	D	E	D	C	G	I	D

Table 11
Correlations between placements across the progressions - 2016

	Engagement	Object Counting	Emotional Literacy	Grip and Manipulation	Crossing Midline	Following Directions	Letter Naming
Object Counting	.522						
Emotional Literacy	.535	.571					
Grip and Manipulation	.510	.500	.527				
Crossing Midline	.427	.386	.402	.520			
Following Directions	.557	.584	.610	.540	.414		
Letter Naming	.465	.598	.489	.492	.383	.580	
Hand Dominance	.410	.414	.453	.581	.476	.443	.341

Table 12
Item difficulty by subgroup - 2016

Progression (scale points)	Index	Total Sample	Female	Male	African American	Hispanic	White
Engagement (0-5)	Classical Item Difficulty	2.90	3.02	2.78	2.67	2.81	3.05
Emotional Literacy (0-10)		4.82	4.93	4.63	4.42	4.41	5.12
Grip and Manipulation (0-6)		3.82	3.96	3.65	3.63	3.75	3.91
Hand Dominance (0-5)		3.16	3.23	3.09	3.07	3.09	3.20
Crossing Midline (0-4)		2.70	2.74	2.64	2.63	2.65	2.74
Object Counting (0-9)		4.55	4.53	4.42	4.10	3.99	4.86
Letter Naming (0-12)		7.96	7.96	7.68	7.41	6.98	8.36
Following Directions (0-12)		6.83	6.95	6.52	6.25	6.19	7.21
Engagement (0-5)	Rasch Item Difficulty	0.05	-0.08	0.01	0.14	-0.07	-0.02
Emotional Literacy (0-10)		0.57	0.57	0.57	0.57	0.57	0.57
Grip and Manipulation (0-6)		-0.27	-0.38	-0.20	-0.27	-0.44	-0.21
Hand Dominance (0-5)		-0.51	-0.59	-0.54	-0.62	-0.70	-0.48
Crossing Midline (0-4)		-0.52	-0.61	-0.48	-0.57	-0.66	-0.46
Object Counting (0-9)		0.50	0.56	0.46	0.53	0.56	0.48
Letter Naming (0-12)		0.01	0.06	0.01	-0.03	0.09	0.01
Following Directions (0-12)		0.17	0.17	0.17	0.17	0.17	0.17

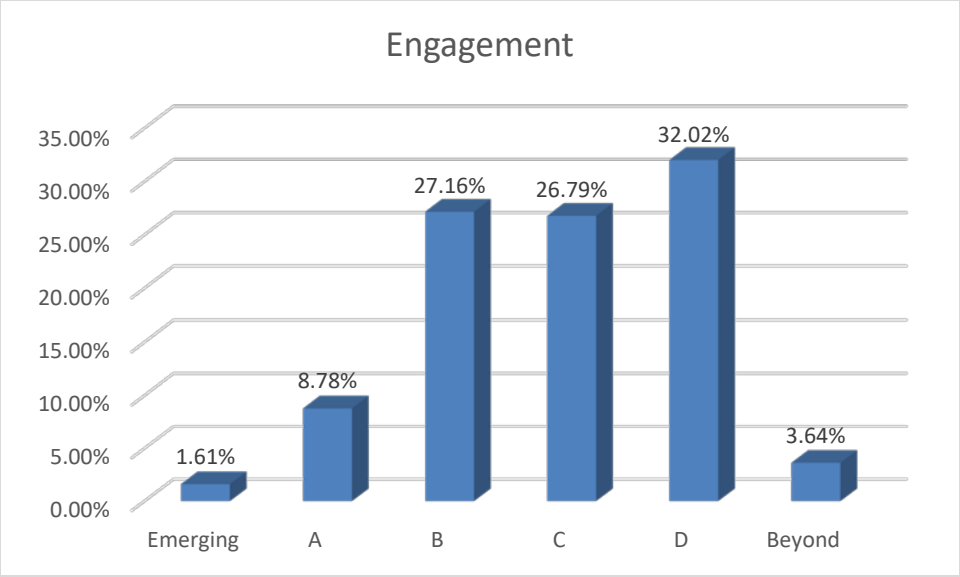

Figure 2. Distribution of progression placements for Engagement – 2016.

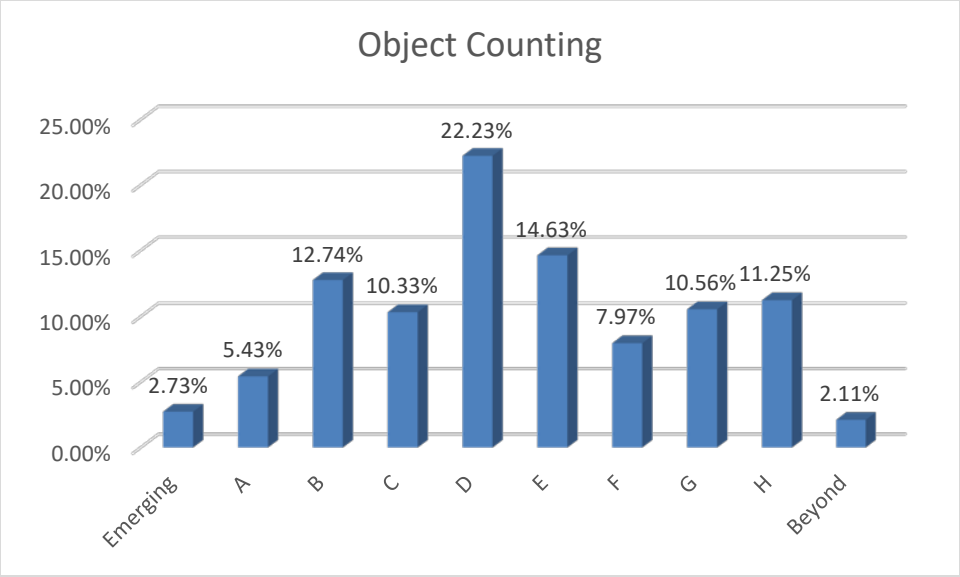

Figure 3. Distribution of progression placements for Object Counting – 2016.

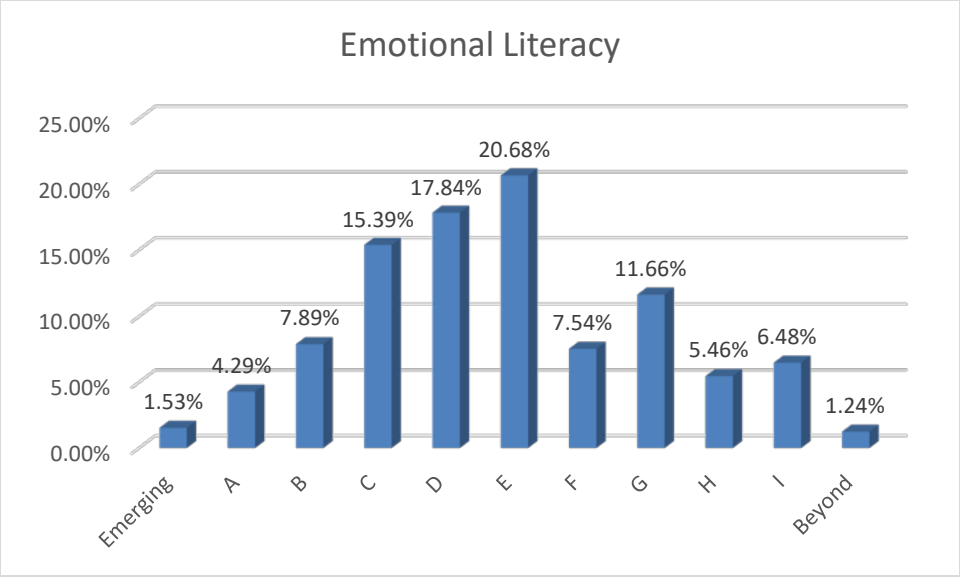

Figure 4. Distribution of progression placements for Emotional Literacy – 2016.

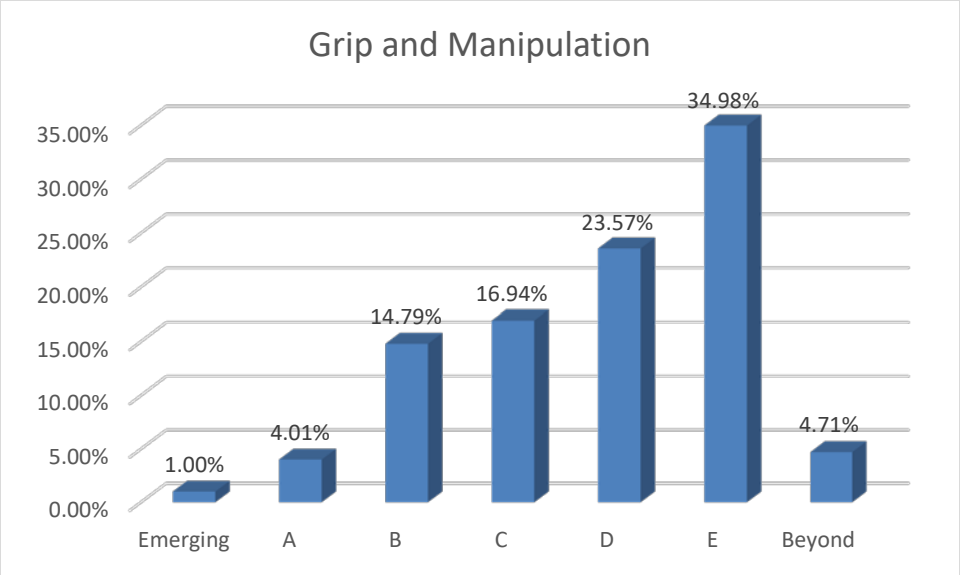

Figure 5. Distribution of progression placements for Grip and Manipulation – 2016.

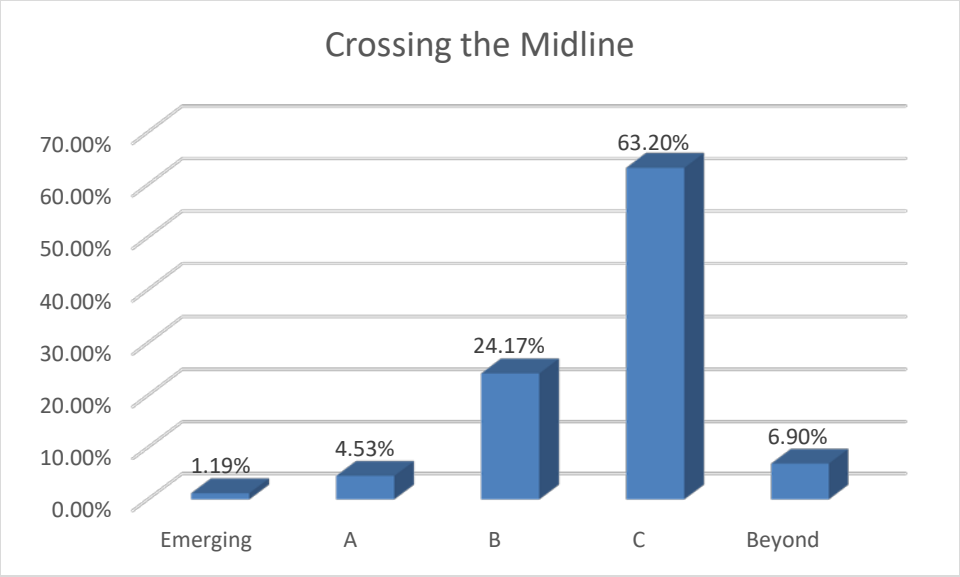

Figure 6. Distribution of progression placements for Crossing the Midline – 2016.

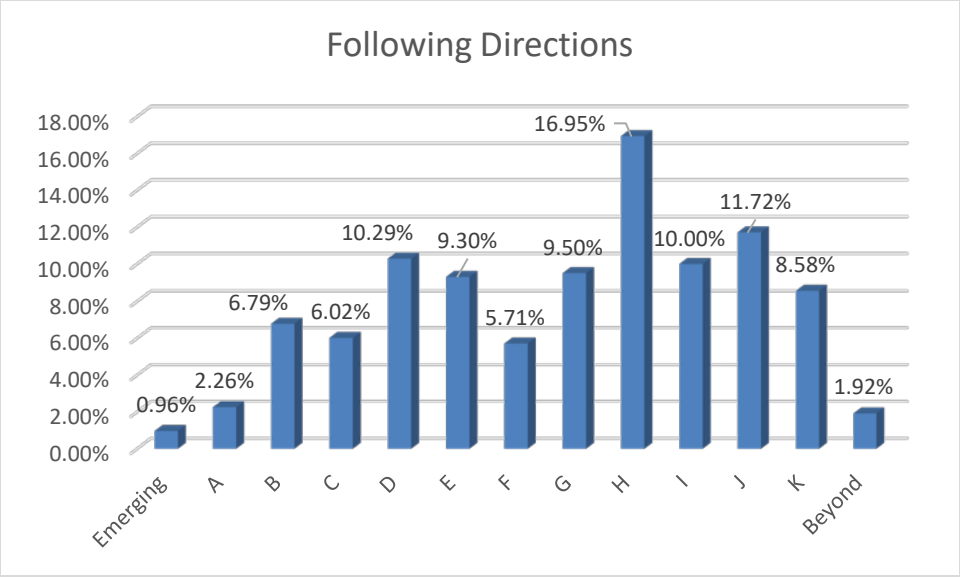

Figure 7. Distribution of progression placements for Following Directions – 2016.

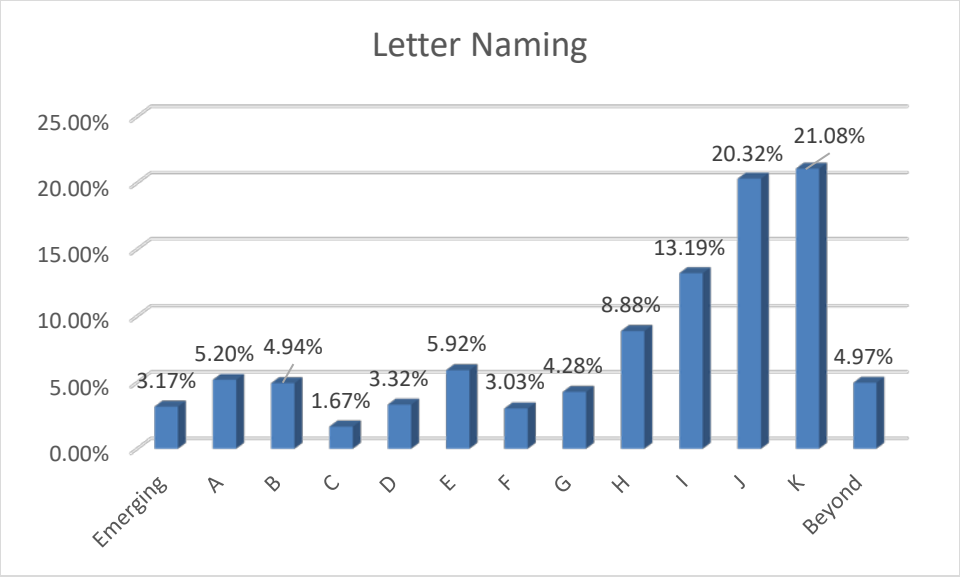

Figure 8. Distribution of progression placements for Letter Naming – 2016.

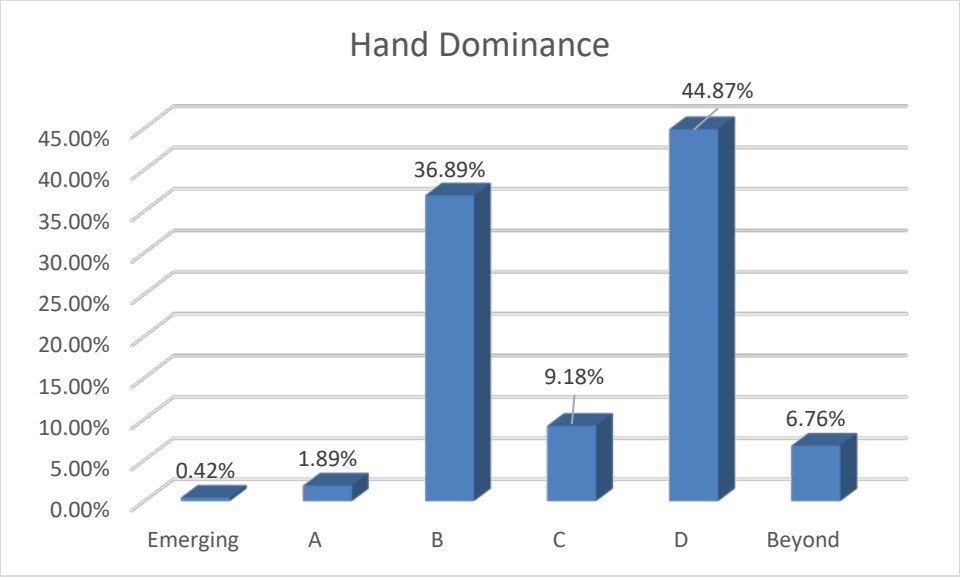

Figure 9. Distribution of progression placements for Hand Dominance – 2016.

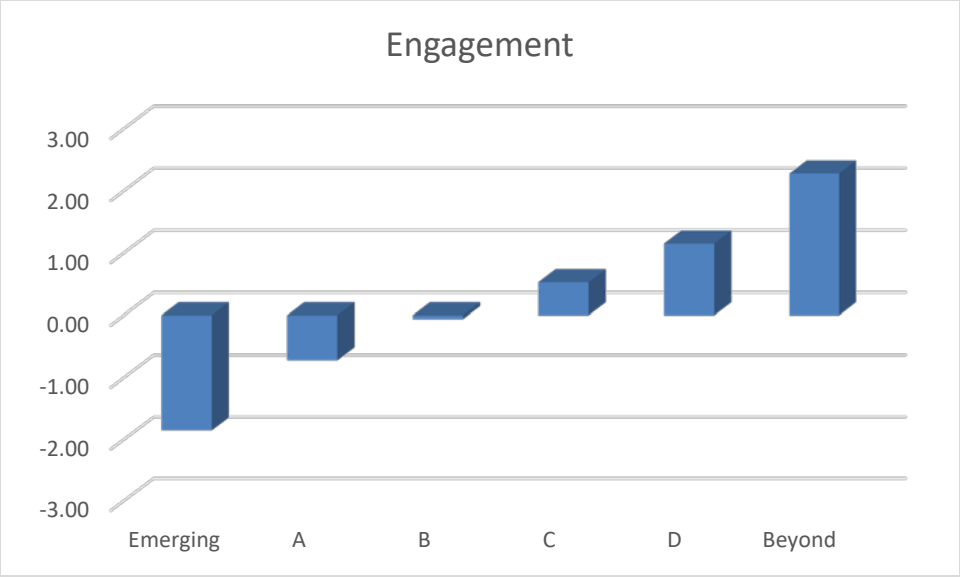

Figure 10. Average Measure Scores by Category for Engagement – 2016.

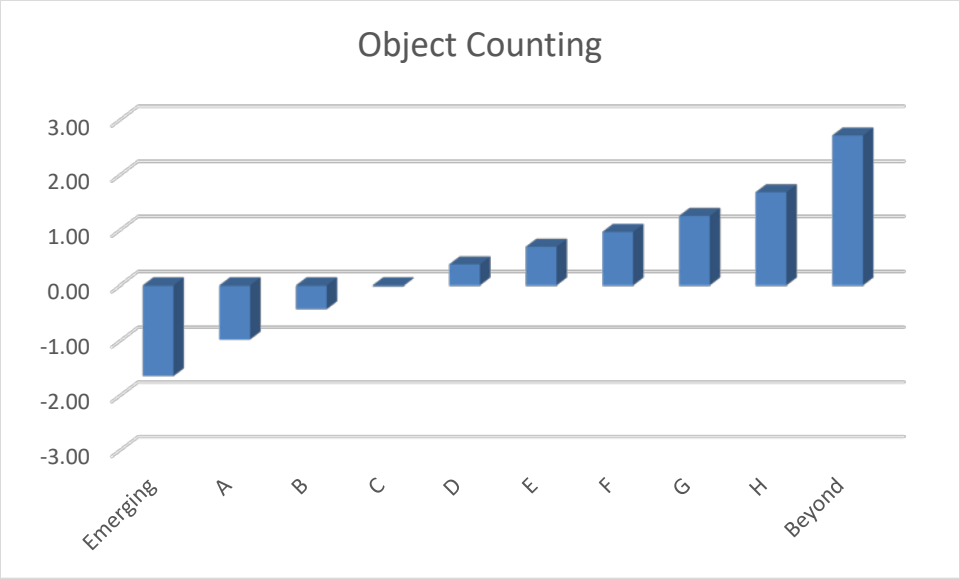

Figure 11. Average Measure Scores by Category for Object Counting – 2016.

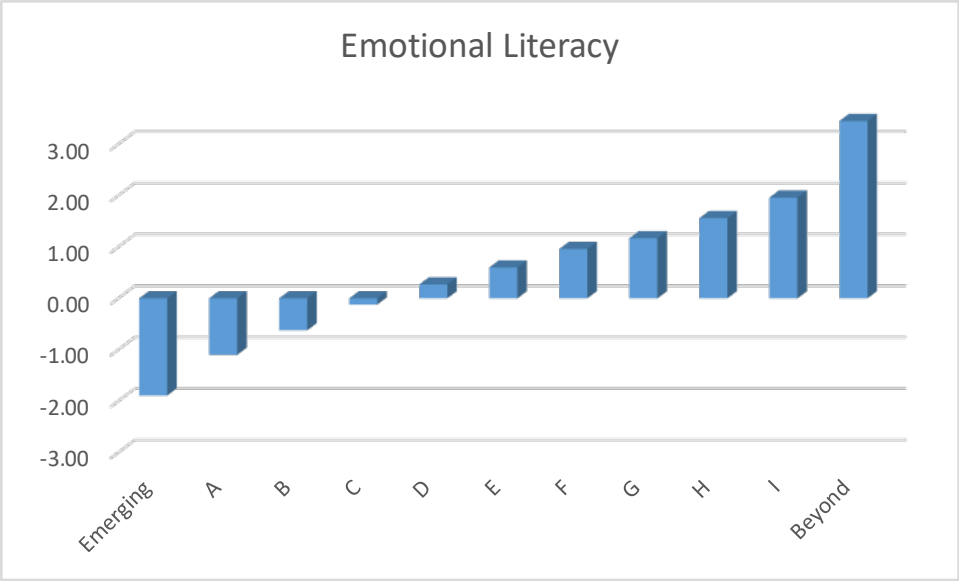

Figure 12. Average Measure Scores by Category for Emotional Literacy – 2016.

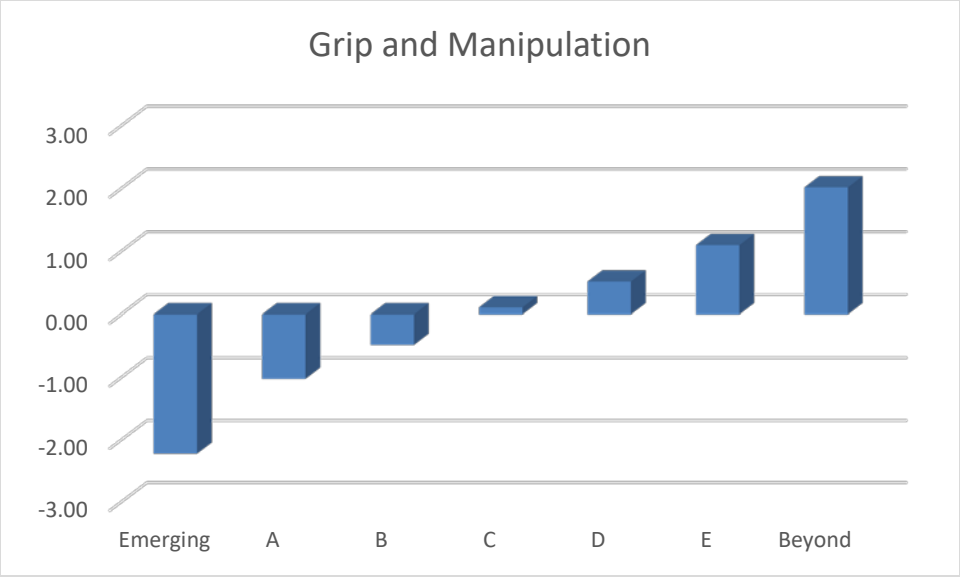

Figure 13. Average Measure Scores by Category for Grip and Manipulation – 2016.

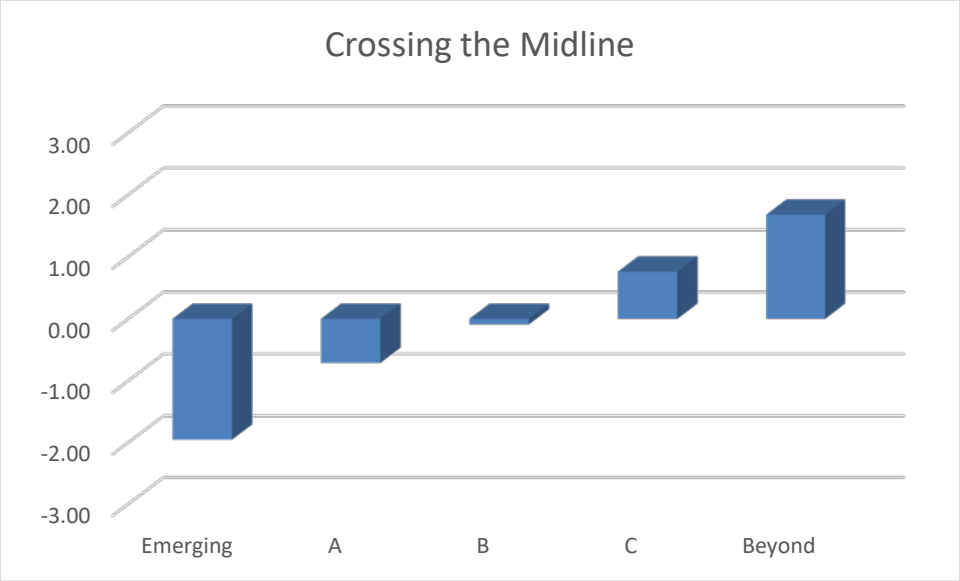

Figure 14. Average Measure Scores by Category for Crossing the Midline – 2016.

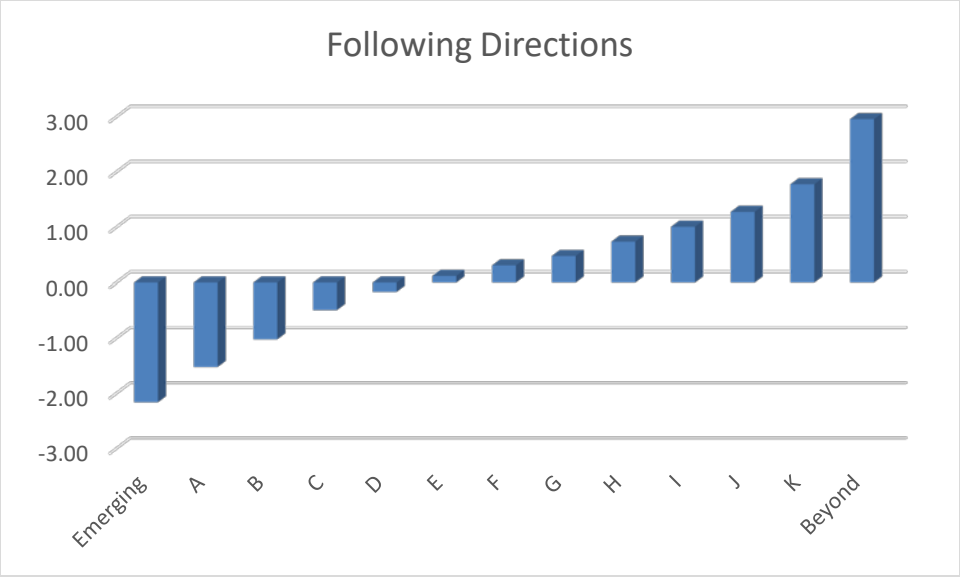

Figure 15. Average Measure Scores by Category for Following Directions – 2016.

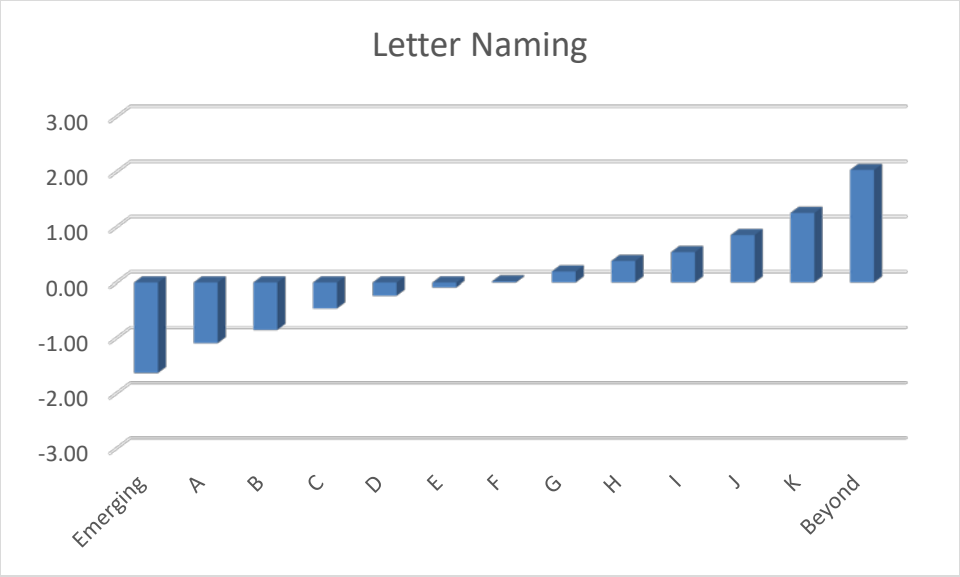

Figure 16. Average Measure Scores by Category for Letter Naming – 2016.

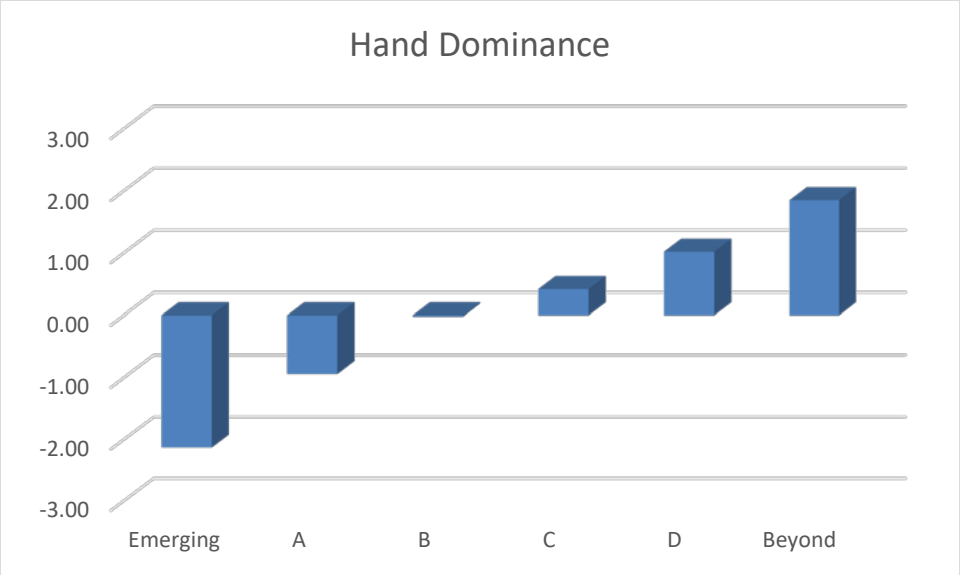

Figure 17. Average Measure Scores by Category for Hand Dominance – 2016.

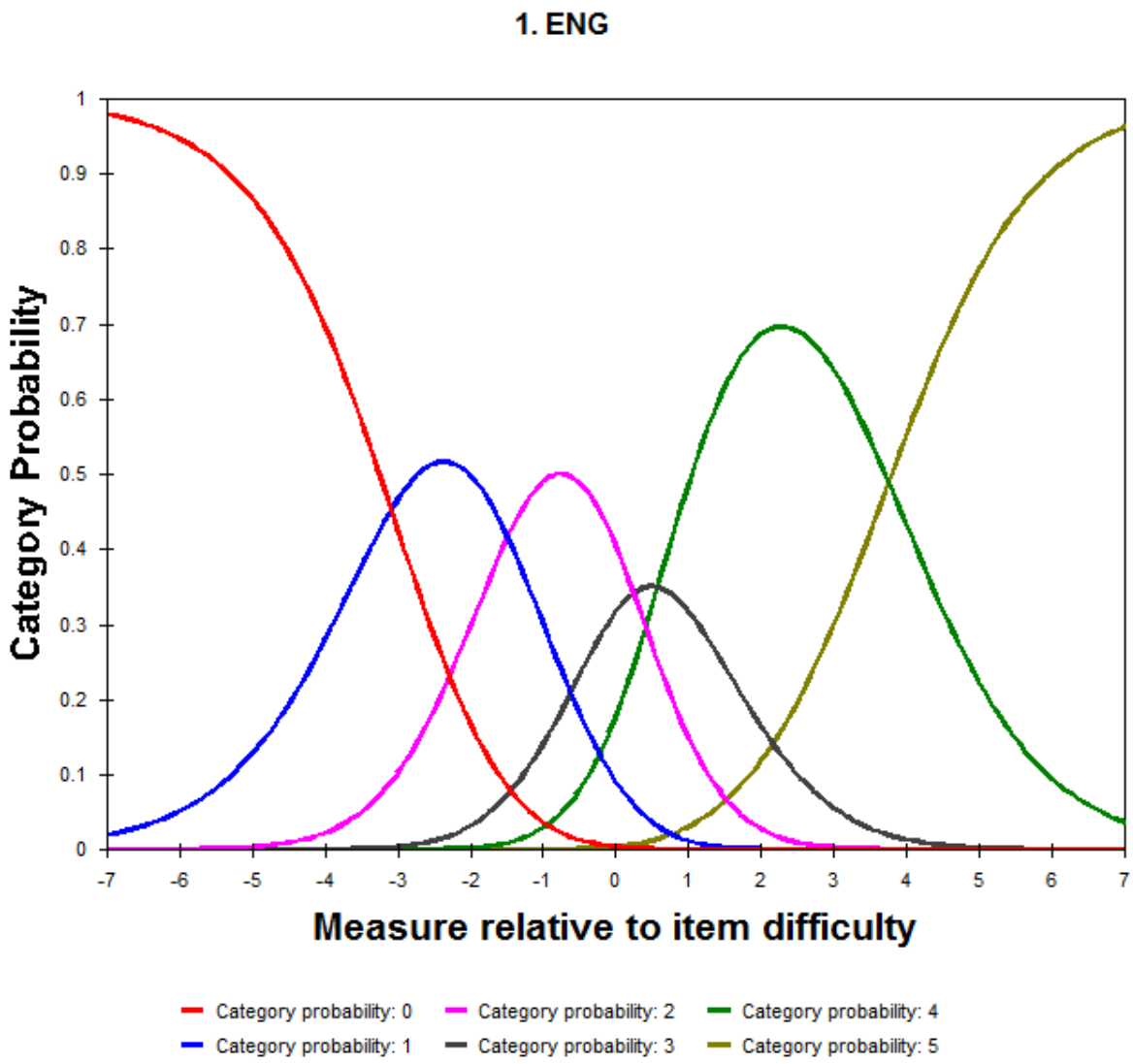

Figure 18. Category Probability Plot for Engagement – 2016.

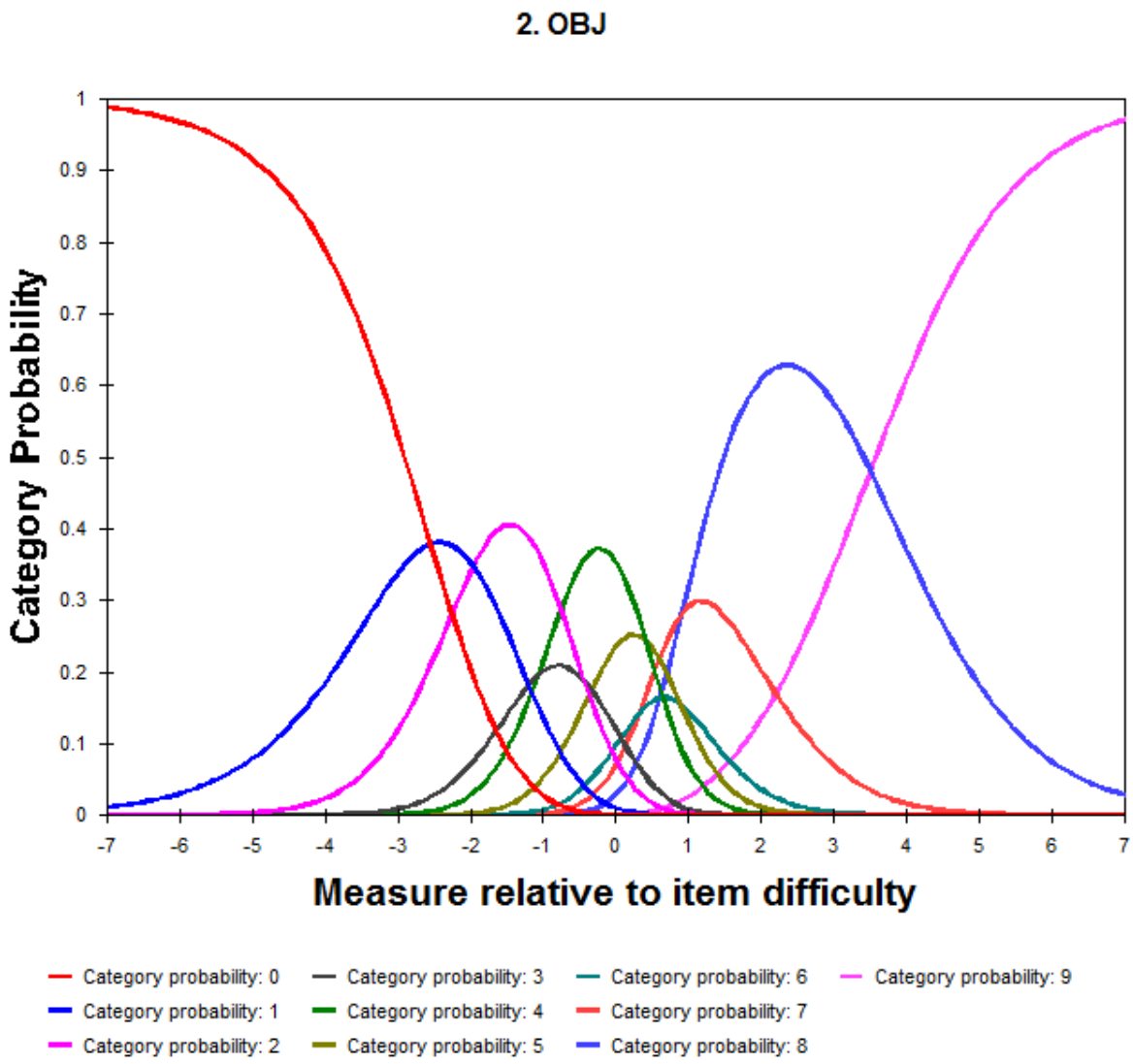

Figure 19. Category Probability Plot for Object Counting – 2016.

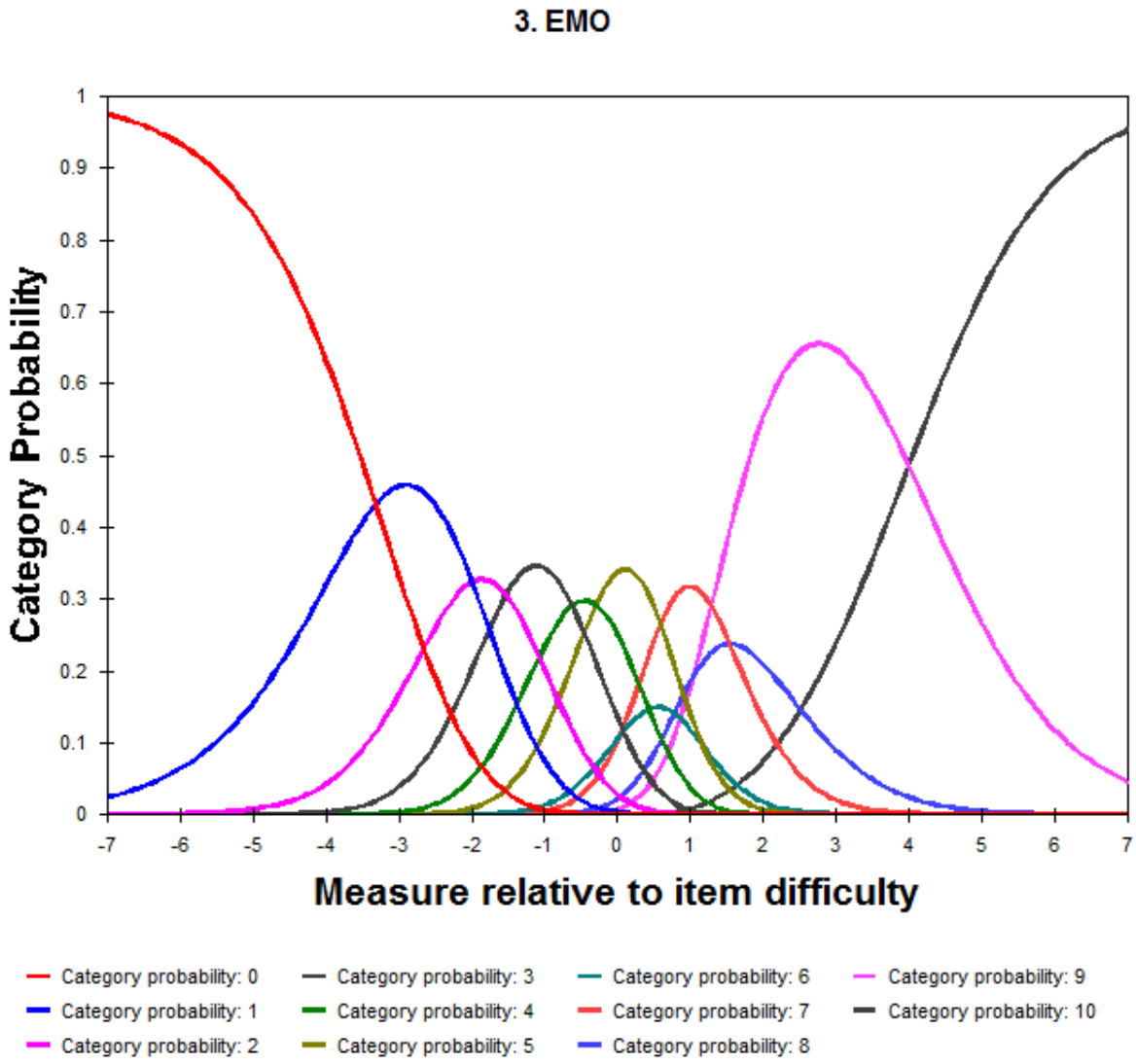

Figure 20. Category Probability Plot for Emotional Literacy – 2016.

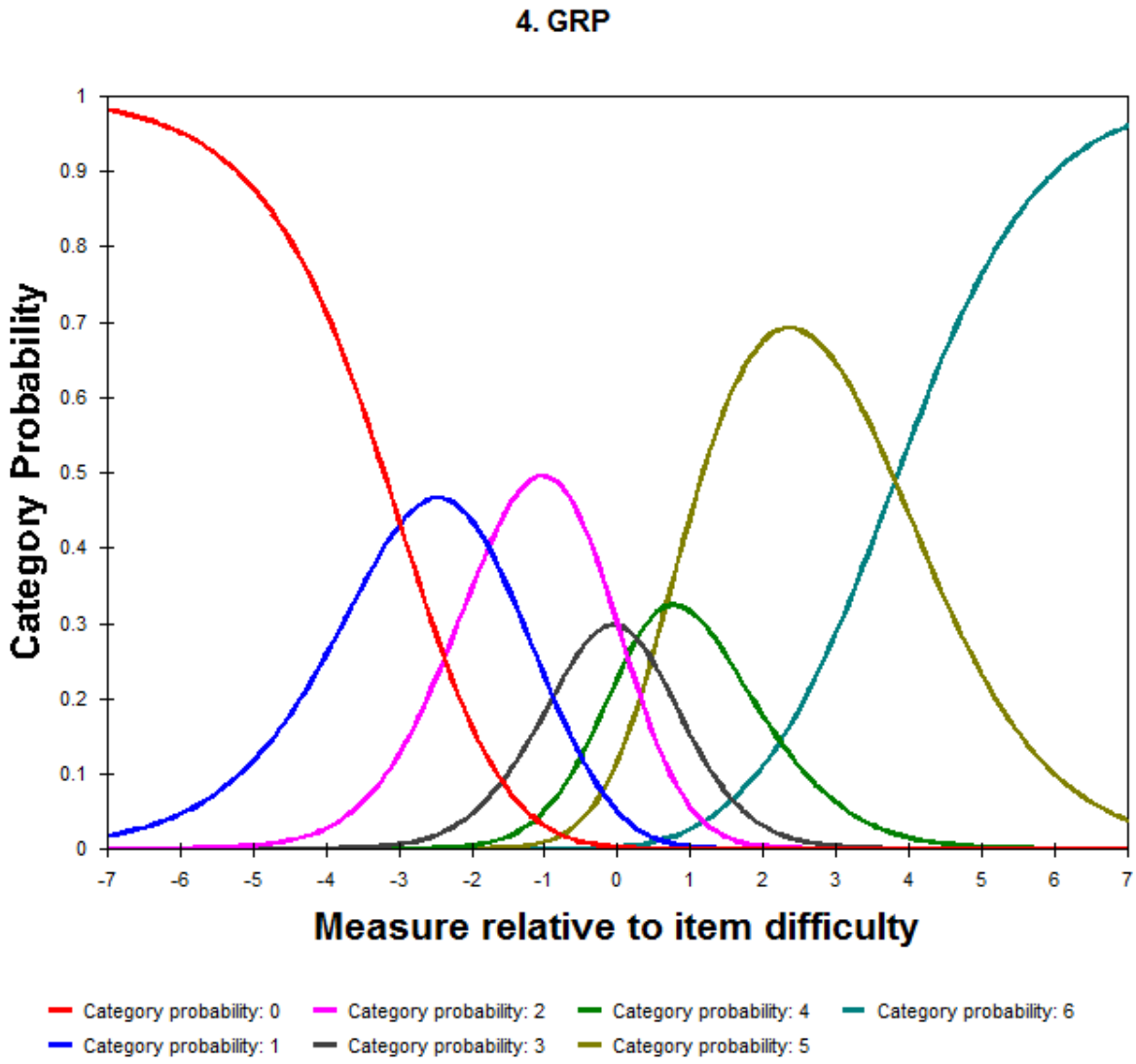

Figure 21. Category Probability Plot for Grip and Manipulation – 2016.

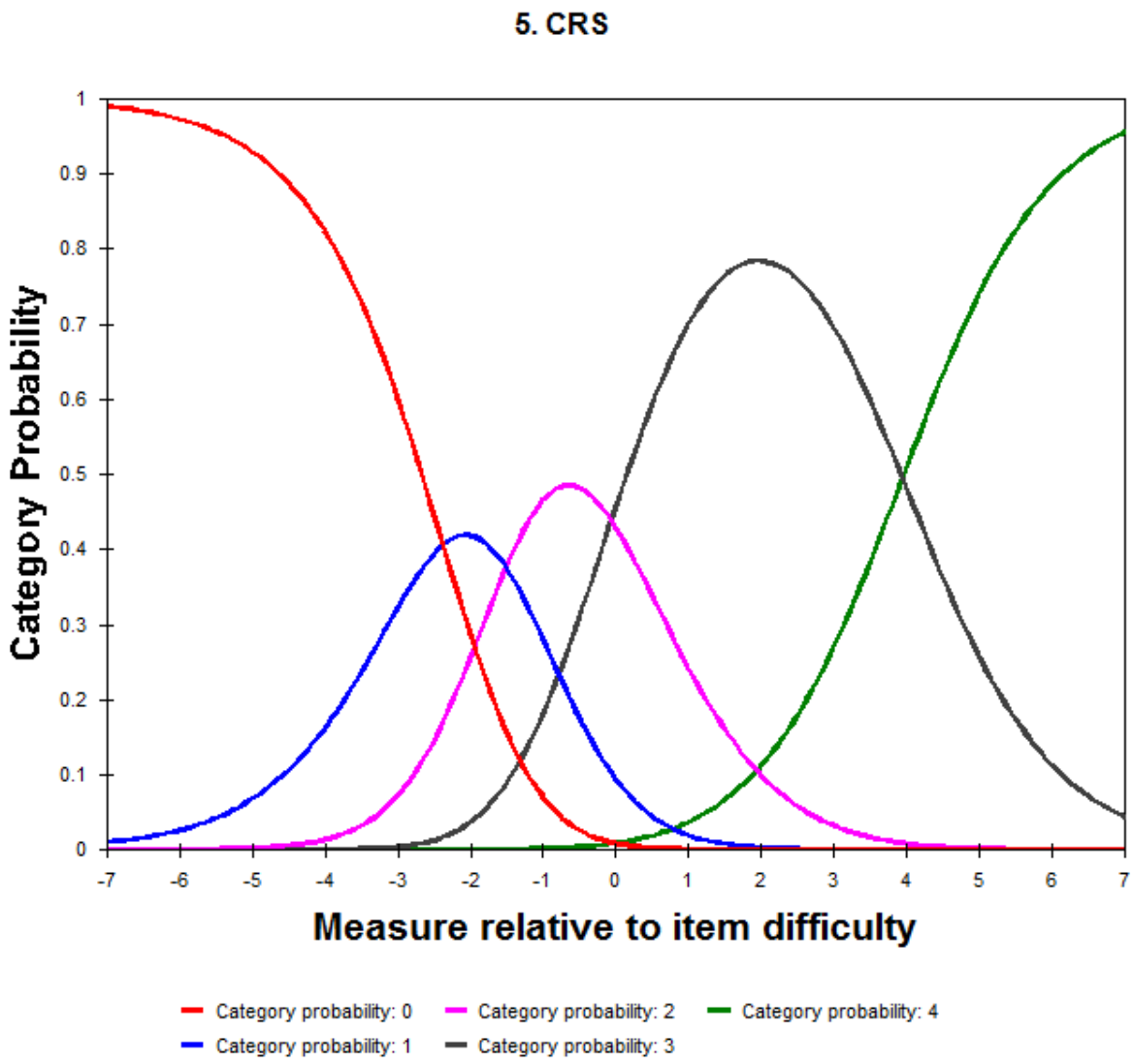

Figure 22. Category Probability Plot for Crossing the Midline – 2016.

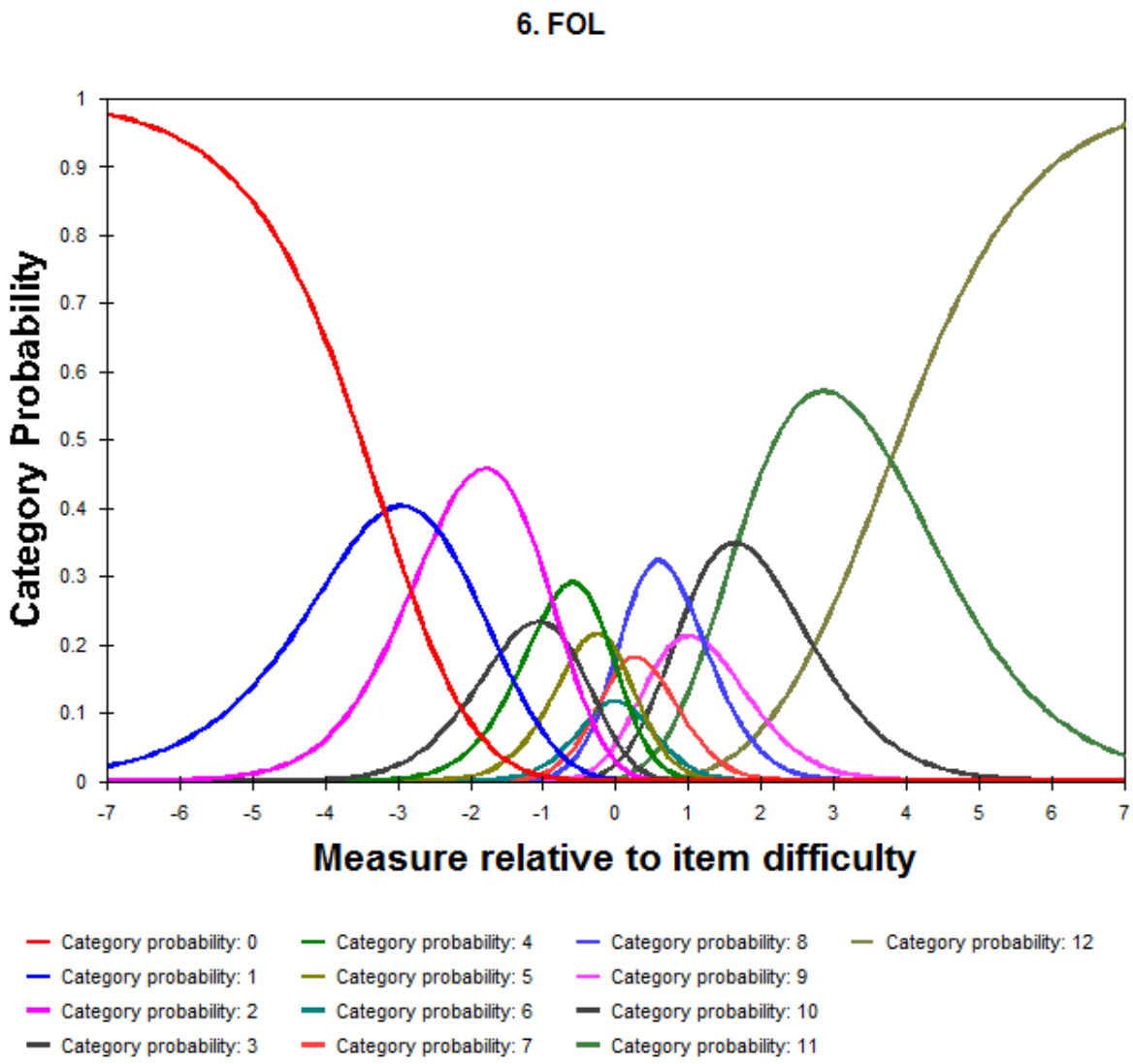

Figure 23. Category Probability Plot for Following Directions – 2016.

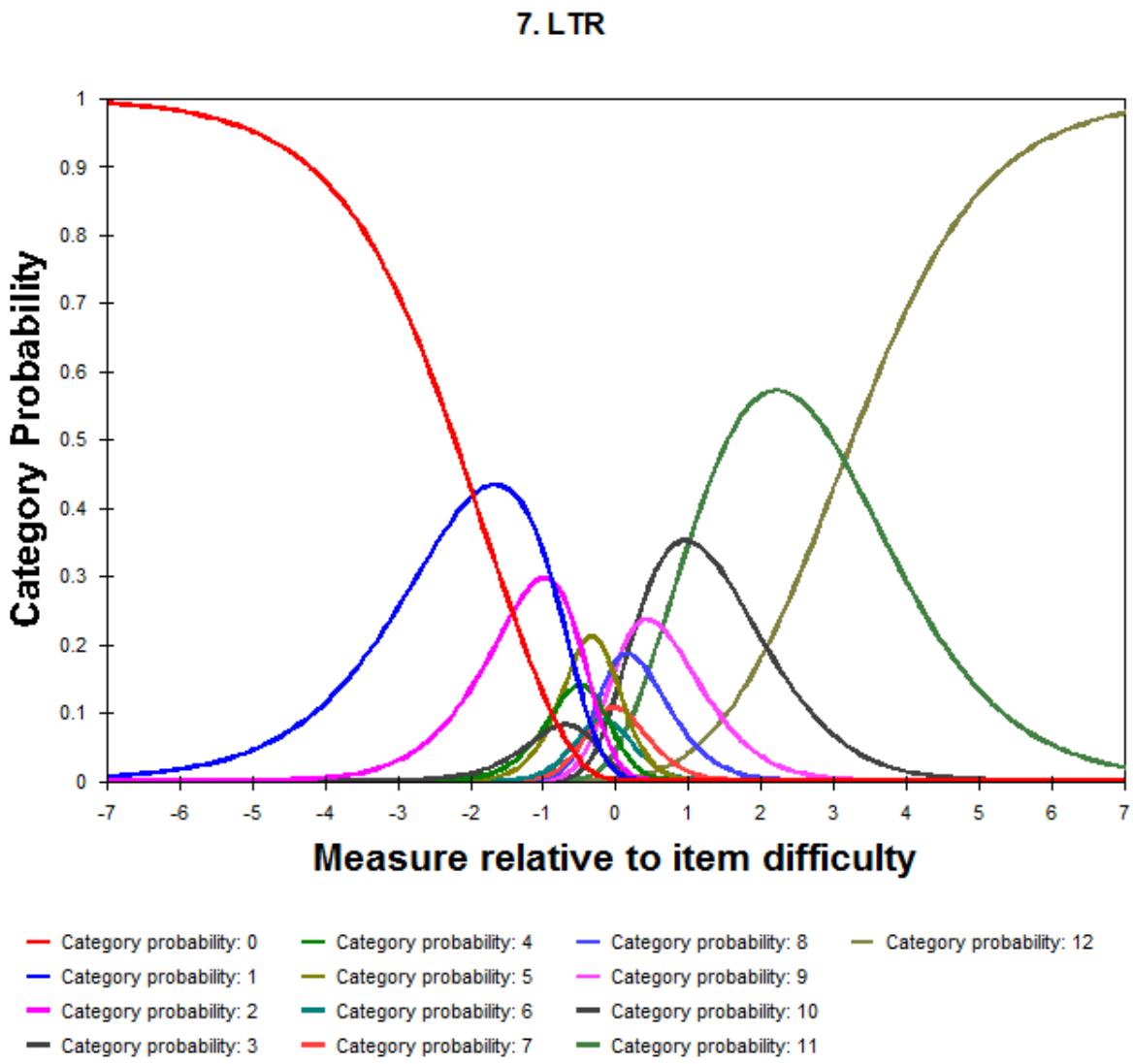

Figure 24. Category Probability Plot for Letter Naming – 2016.

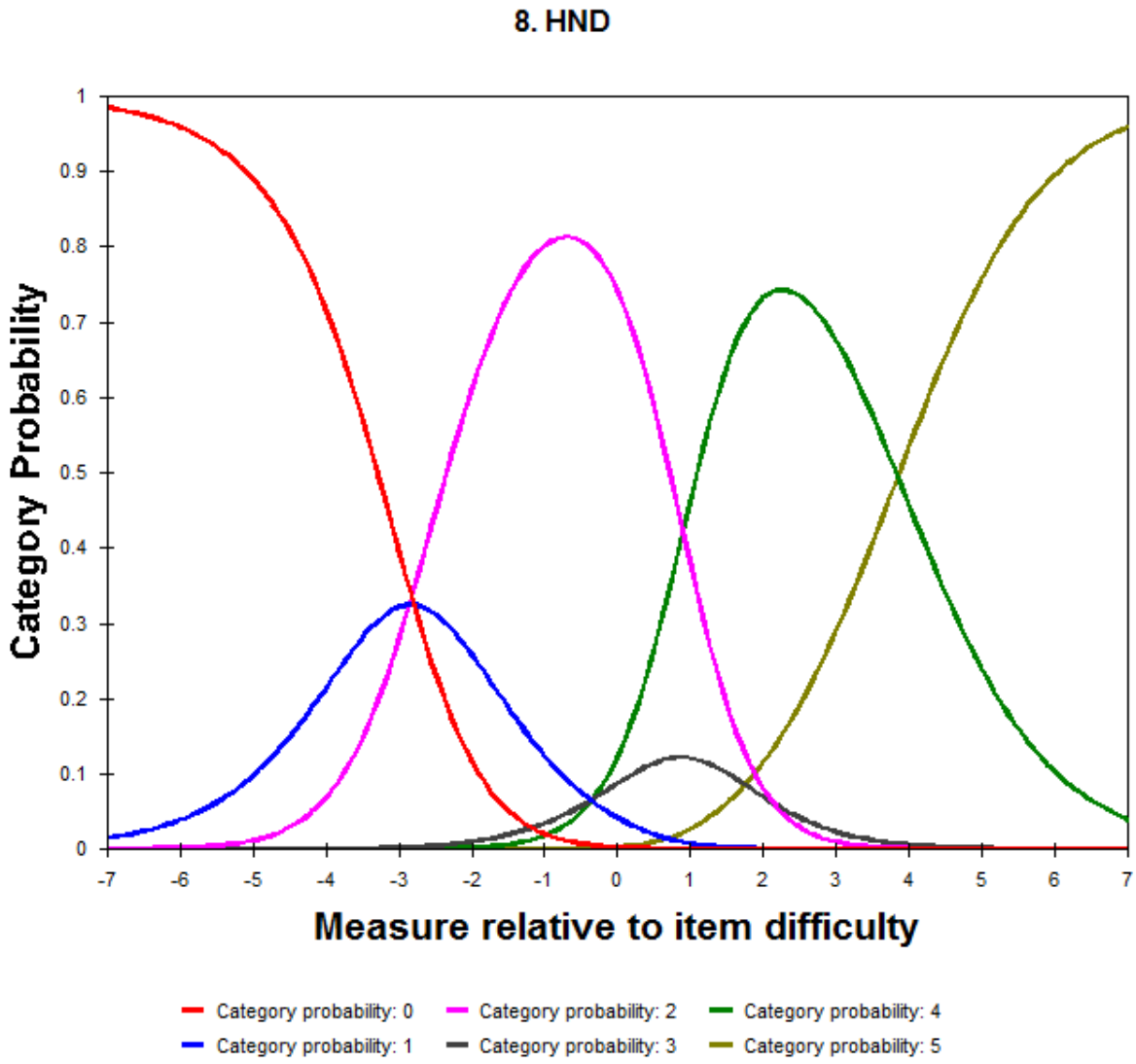

Figure 25. Category Probability Plot for Hand Dominance – 2016.

Results from the fall, 2017 Assessment

Dimensionality

Rasch modeling assumes what is called unidimensionality, meaning that the progressions in question measure one and only one underlying latent construct. In the case of the KEA, this latent construct might be considered global development of the whole child. The unidimensionality of the total score, or scale, was evaluated by using Mean Square (MNSQ) progression fit statistics and Rasch Principal Components Analysis of residuals (PCAR). The MNSQ fit values between 0.6 and 1.4 are considered reasonable for rating scale progressions (Bond & Fox, 2007). MNSQ values less than 2.0 can indicate that a progression, though not fitting optimally with the measurement model, can still contribute useful information to the overall score on the measure. Progressions with mean square values of between 1.4 and 2.0 can be considered potentially unproductive for the construction of measurement scales, but not degrading to the quality of the information provided by the scale (Linacre, 2002). Infit statistics indicate the fit of individual progression response patterns to the measurement model. They also address the possibility of secondary dimensions and fit to the underlying construct. Outfit statistics are sensitive to outliers; that is responses that show great differences between person responses and progression difficulties. They are also sensitive to unusual and unexpected progression response patterns.

For PCAR, a variance of greater than 50% explained by measures is considered good, and offers support for scale unidimensionality. If a secondary dimension has an eigenvalue of smaller than 3 and accounts for less than approximately 5% of the unexplained variance, unidimensionality is considered plausible (Linacre, 2012).

The PCAR showed that the Rasch dimension explained the majority of the variance in the data (67.1%) with an eigenvalue of 16.3, relative to the total eigenvalue of 24.3. The first contrast (the largest potential secondary dimension) had an eigenvalue of 1.6 and accounted for 6.8% of the

unexplained variance. When the first contrast was examined further, there was some evidence that Letter Naming and Object Counting might comprise a possible second factor, with the progressions focused on physical and social development comprising the first factor. This finding was very similar to what was found with the 2016 data. However, this evidence was weak and needs to be monitored in the future as more progressions are implemented.

The fit statistics for all of the progressions were well within acceptable limits (see Table 1). The infit MNSQ values ranged from 0.88 to 1.27. The outfit MNSQ values ranged from 0.88 to 1.32. The progression to total score correlations, with each progression excluded from the total score, ranged from .57 to .78. The progression to total score correlations, with each progression included in the total score, ranged from .58 to .78. In summary, these model fit statistics when taken together generally suggest that the data does in fact fit the Rasch PCM very well. These results also indicated that the data satisfied the unidimensionality assumption of the Rasch model.

Item Difficulty Measures

The progression location hierarchy appeared to be generally consistent with the expected developmental trajectory for typically developing kindergarten children. Table 1 lists the progression difficulty estimates from highest to lowest along with the standard errors for these estimates and the associated fit statistics. These results were evaluated using the final data available at the end of the fall 60-day KEA assessment time period. The progressions pertaining to a child's ability to cross the midline and demonstrate hand dominance were estimated as the relatively easiest progressions (-.45, -.56). The progressions pertaining to a child's ability to follow directions, engage in self-selected activities, name letters, and demonstrate grip and manipulation skills were found to be of average difficulty level (-.21 to .18). The progressions pertaining to object counting and a child's ability to demonstrate emotional literacy were to be the most difficult (.41, .54).

The range of progression difficulties (-.56 to .54) was found to be relatively narrow and it will be ideal to add progressions with a wider range of difficulty levels in the future. This can be seen in Figure 1. This figure displays the Item Person Map. On the left side of the center of the map, the distribution of total scores for the population of children is displayed. This distribution conforms closely to a unimodal and symmetrical shape and indicates that the total measure score is functioning well to spread children out according to underlying overall developmental status. The right side of the map indicates the location of each progression. The progression locations, or difficulty estimates, indicate that the progressions are functioning well to separate children near the center of the distribution and are less useful for spreading out children at the upper and extreme lower ends of the distribution. The practical implication for teachers is that these eight progressions may be relatively less useful for understanding and supporting the developmental progress of children with very low or more highly developed global development across domains.

When the progression rating scale anchor point, or category, locations are considered, these values come closer to matching the range of abilities of the children assessed. In tables 2 through 9, the Andrich thresholds are reported. These values indicate the ability locations that form the model estimated boundaries between the rating scale or progression categories. These locations indicate where on the underlying ability scale, or total score, the probability becomes higher that a child will be placed at the next highest category on the progression, relative to the previous anchor point. The values were as follows: Engagement = -3.00 – 3.76, Object Counting = -2.97 – 3.56, Emotional Literacy = -3.30 – 3.98, Grip and Manipulation = -3.11 – 3.85, Crossing the Midline = -2.52 – 4.01, Following Directions = -3.22 – 3.89, Letter Naming = -1.99 – 3.36, and Hand Dominance = -3.09 – 3.80. These values more closely match the full range of ability estimates on the total score and provide reasonable separation of children according to underlying ability.

In summary, the developmental pathway that is formed indicates a pathway from the easiest to the most difficult progressions that generally aligns with expectations from developmental theory. It is also important to recognize, as indicated, that the range of progression difficulties is effectively much wider than the results indicate when considering the separation created between children by the range of rating scale anchor point threshold locations.

Reliability

Reliability was evaluated using the following Rasch indexes: the person separation index, item separation index, person reliability, and item reliability. Item (progression) and person reliabilities were evaluated using both sample-based and model-based coefficients. The person separation index, an estimate of the adjusted person standard deviation divided by the average measurement error, indicates how well the instrument can discriminate persons on each of the constructs. The item (progression) separation index indicates an estimate in standard error units of the spread or separation of progressions along the measurement constructs. Reliability separation indexes greater than 2 are considered adequate, and indexes greater than 3 are considered high (Bond & Fox, 2007). High person or item (progression) reliability means that there is a high probability of replicating the same separation of persons or progressions across measurements. Specifically, person separation reliability estimates the replicability of person placement across other progressions measuring the same construct. Similarly, progression separation reliability estimates the replicability of progression placement along the construct developmental pathway if the same progressions were given to another sample with similar ability levels. The person reliability provided is similar to the classical or traditional test reliability whereas the progression reliability has no classical equivalent. Low values in person and progression reliability may indicate a narrow range of person or progression measures. It may also indicate that the number of progressions or the sample

size under study is too small for stable estimates (Linacre, 2009). Reliability was also evaluated using Cronbach's alpha measure of internal consistency.

The item (progression) reliability values, both sample-based and model-based, were greater than .99. The item (progression) separation indexes were also very high: sample-based = 110.9 and model-based = 111.6. Taken together, these findings indicate it is reasonable to expect highly consistent estimates of progression difficulty levels across samples. The sample-based person separation index was 2.76 and the model-based value was 3.16. The sample-based person reliability index was .88 and the model-based value was .91. The Cronbach's alpha value for the total score was .86. Based on these reliability indexes, the total scores appear to yield adequately reliable information from this sample. Specifically, these results indicate that it is reasonable to expect reliable estimates of child overall ability levels when teachers use the KEA to place kindergarten children along the developmental progressions, and those individual progression scores are transformed into a composite or total score. It is important to note that these results address reliability issues related to the use of a total score and may be very different from the results of an inter-rater reliability study.

Rating Scale Category Effectiveness

A rating scale with demonstrated category effectiveness yields evidence that raters are using the scale as it was intended to be used. This means that raters can use the scale to discriminate between responses with true underlying differences on the construct being measured. In the case of the KEA, rating scale category effectiveness is a measure of the validity of the data elicited by the developmental progressions. Developmental progressions with effective rating scales yield valid data that can be used to place children along a continuum of development so that the placements both reflect the true developmental status of each child and can be used by teachers to differentiate instruction and support growth, learning, and development. As with the 2016 data, this study

focused on the following research questions in an effort to begin to understand the rating scale category effectiveness of the KEA progressions as used by North Carolina kindergarten teachers:

1.) What are the characteristics of the distributions of placements on each of the progressions?
2.) Do the mean total scores of the children placed in each category increase monotonically along the rating scale for each of the progressions?
3.) Do the thresholds between rating scale categories increase monotonically along the rating scale for each of the progressions?
4.) Do the category probability plots indicate distinct probability distributions for each rating scale point for each of the progressions?

To address research question 1, the center, shape, and spread of the distribution of ratings was examined for each progression. It is recommended that for each progression, each rating scale category needs to be assigned to a minimum of 10 children. All rating scale categories should be used by the raters and each category should be assigned to enough children to allow for reasonable statistical estimates within the Rasch modeling process. These criteria were easily met for all eight progressions. Across the eight progressions, the full range of categories, from "Emerging" to "Beyond", was used by the teachers. Tables 2 through 9 include the number and percent of children assigned to each rating scale category. Table 10 includes the mean, median, and standard deviation for each progression. The median is also reported as the median lettered category for each progression.

Figures 2 through 9 display the shape of the distribution of ratings for each progression through a simple bar chart of the percentage of children placed in each rating scale category. These charts indicate a reasonably unimodal and symmetrical shape to the distribution of ratings for each progression with several notable exceptions. For Engagement, there were relatively few placements in the extreme (lowest or "Emerging" and highest or "Beyond") categories. For Object Counting,

the extreme categories very also used relatively infrequently as was category F. For Emotional Literacy, the lowest and highest categories were used relatively infrequently as was category F. For Grip and Manipulation, the distribution was negatively skewed, and the lowest category was relatively infrequently used. For Crossing the Midline, the distribution was also negatively skewed, and the lowest category was relatively infrequently used. For Following Directions, both the lowest and highest category, and category F were relatively infrequently used. For Letter Naming, the distribution was negatively skewed, and the lowest category was relatively infrequently used as were categories C, D, and F. For Hand Dominance, both the lowest category and category C were relatively infrequently used.

To address research question 2, the average of the overall ability estimates, based on the total progression scores, for all children in the sample who were placed at a particular response category or scale point on each of the developmental progressions was examined. Average measure scores should advance monotonically with rating scale category values (Bond & Fox, 2007). Tables 2 through 9, under the column labeled Observed Average, demonstrate that the average total scores did increase as expected across all rating scale categories for each of the progressions. This finding is a very positive result for the validity of the progressions and is also illustrated graphically in figures 10 through 17.

To address research question 3, the category thresholds were examined. Thresholds (also called step calibrations) are the difficulty levels estimated as the point on the total score at which teachers are more likely to choose one response category or rating scale point over the previous step on the progression (Bond & Fox, 2007). For this study the Andrich thresholds from the Partial Credit Model were used. Thresholds should also increase monotonically along the rating scale categories. These values are reported in tables 2 through 9 under the column labeled Andrich Threshold. For three of the progressions, Engagement, Grip and Manipulation, and Crossing the

Midline, all thresholds increased monotonically as expected. However, the remaining five progressions all had at least one disordered threshold. These are indicated by boxes around italicized threshold values. For Object Counting, categories D and G had disordered thresholds. For Emotional Literacy, categories G and I had disordered thresholds. For Following Directions, categories D, F, and J had disordered thresholds. For Letter Naming, categories C, D, F, and H had disordered thresholds. For Hand Dominance, category D had a disordered threshold. Each of these identified categories should be examined further as disordered thresholds present a threat to the validity of the rating scale data and any inferences made from the data.

To address research question 4, category probability plots were examined. These plots indicate the probability distribution for a child being placed on a particular response category, or level on each developmental progression, given their overall ability or total measure score. The plots should contain distinct and minimally overlapping probability distributions for each rating scale category. The magnitude of the distances between adjacent category thresholds should be large enough so that each step defines a distinct position and each category has a distinct peak in the category probability curve plot (Bond & Fox, 2007). Figures 18 through 25 displays these plots. Overall, these plots suggest substantial difficulties with the use of the rating scales. These may suggest a need to reduce or combine categories, improve the definitions of the category anchors, improve teacher training and understanding of the distinct differences between categories, a need for teachers to collect higher quality evidences, or some combination of these factors. For Engagement, category C shows substantial overlap with adjacent categories. For Object Counting, categories C through G show substantial overlap with adjacent categories. For Emotional Literacy, categories D through H show substantial overlap with adjacent categories. For Grip and Manipulation, categories C and D show substantial overlap with adjacent categories. For Crossing the Midline the plot indicates appropriately distinct probability distributions. For Following Directions, categories C

through J show substantial overlap with adjacent categories. For Letter Naming, categories C through J show substantial overlap with adjacent categories. For Hand Dominance categories A and D show substantial overlap with adjacent categories.

Differences by Subgroup

Another type of evidence for the validity of the information produced by developmental rating scales is extent to which different subgroups of children receive similar scores. Specifically, two children with the same underlying ability should receive the same placement on each developmental progression, and this expectation should be sustained independent of subgroup membership. If children with the same underlying ability receive different placements on the progressions and those differences are systematic based on subgroup membership, then the possibility exists for some level of bias to be inherent in the assessment process. This bias could be related to item content, rater biases, training, or other factors. However, it is unacceptable under any conditions of use. Therefore, subgroup differences based on both gender and race / ethnicity were examined using both classical and modern measurement strategies.

Classical item difficulty was examined by observing the mean score on each progression for the total sample and for the subgroups of interest. Table 12 displays these values. There were no substantial differences between subgroups based on either gender or race / ethnicity. There were only a few exceptions to this finding. White children, on average, tended to be placed somewhat higher (.78 - .91 scale points) than their African American or Hispanic counterparts on the Emotional Literacy and Object Counting progressions. White children, on average, tended to be placed moderately higher (.99 – 1.41 scale points) than their African American or Hispanic counterparts on the Letter Naming and Following Directions progressions. These differences in item difficulty were also examined using Rasch modeling. This method investigates the possibility of Differential Item Functioning by examining the item difficulty estimates by subgroup while

controlling for underlying ability estimates on the total score. This method, therefore, effectively compares children across the subgroups who have the same underlying total ability estimates. There were no substantial differences between item difficulty estimates based on subgroups using this method. Differences in item difficulty estimates greater than or equal to .64 are considered large, .43 - .63 moderate, and less than .43 are considered negligible (Zwick, Thayer, & Lewis, 1999)

The separate item difficulty estimates for males and females are listed in Table 12. The differences between estimates for male and female children, in logit units, ranged from .00 to .19. The separate item difficulty estimates for white, African American, and Hispanic children are also listed in Table 12. The differences between estimates for white and African American children, in logit units, ranged from .00 to .17. The differences between estimates for white and Hispanic children, in logit units, ranged from .00 to .17.

Table 1
Item level statistics and difficulty estimates - 2017

Progression	Item Difficulty	SE	Infit Mnsq	Outfit Mnsq	Item-Measure *r* Item Included	Item Excluded
Emotional Literacy	0.54	< .005	1.01	1.03	0.77	0.77
Object Counting	0.41	< .005	0.96	0.97	0.76	0.75
Following Directions	0.18	< .005	0.99	1.00	0.78	0.78
Engagement	0.07	< .005	0.88	0.88	0.68	0.64
Letter Naming	0.01	< .005	1.27	1.32	0.72	0.75
Grip and Manipulation	-0.21	< .005	0.93	0.92	0.69	0.67
Crossing the Midline	-0.45	0.01	0.97	0.97	0.58	0.57
Hand Dominance	-0.56	< .005	1.01	1.01	0.62	0.62

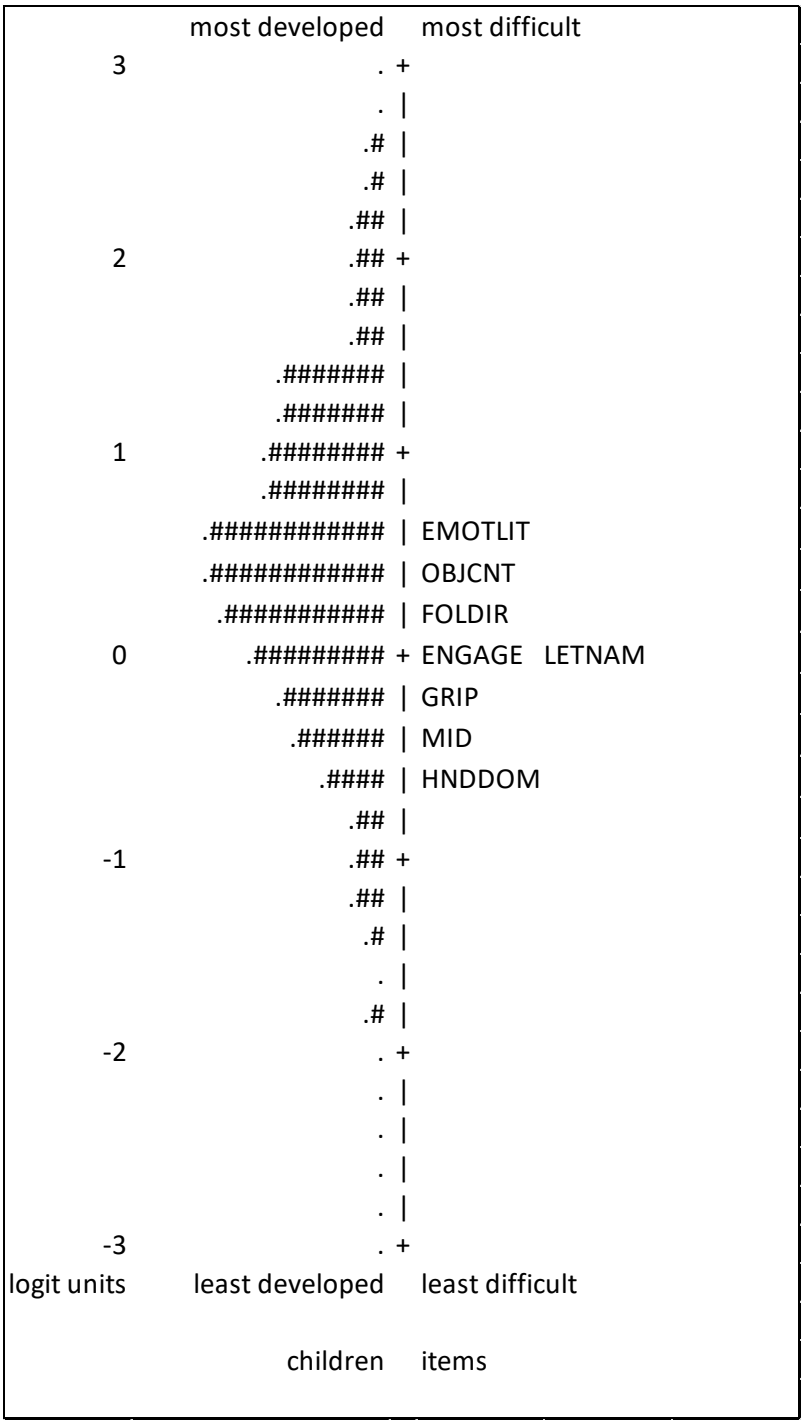

Figure 1. Item Person Map - 2017.

Table 2
Category thresholds and observed average measure scores for Engagement - 2017

Progression Categories	Counts	Percent	Observed Average	Expected	Infit Mnsq	Outfit Mnsq	Andrich Threshold
Emerging	1894	1.85%	-1.76	-1.84	1.08	1.07	-----
A	9228	9.03%	-0.76	-0.68	0.91	0.91	-3.00
B	26990	26.42%	-0.07	0	0.86	0.84	-1.46
C	27706	27.12%	0.51	0.51	0.78	0.76	0.16
D	32573	31.89%	1.14	1.07	0.87	0.88	0.55
Beyond	3754	3.68%	2.3	2.18	0.95	0.96	3.75

Table 3
Category thresholds and observed average measure scores for Object Counting - 2017

Progression Categories	Counts	Percent	Observed Average	Expected	Infit Mnsq	Outfit Mnsq	Andrich Threshold
Emerging	2117	2.08%	-1.94	-1.95	1.15	1.07	-----
A	5868	5.76%	-0.98	-0.99	1.00	0.99	-2.97
B	11614	11.40%	-0.52	-0.40	0.80	0.83	-1.77
C	12862	12.62%	-0.04	-0.01	0.74	0.74	-0.71
D	23478	23.04%	0.36	0.30	0.92	0.90	***-0.86***
E	11115	10.91%	0.67	0.59	0.91	0.94	0.78
F	8514	8.35%	0.91	0.88	1.00	1.07	0.59
G	12817	12.58%	1.20	1.22	1.05	1.08	***0.22***
H	11262	11.05%	1.67	1.70	1.07	1.10	1.15
Beyond	2274	2.23%	2.83	2.82	1.24	1.10	3.56

Table 4
Category thresholds and observed average measure scores for Emotional Literacy - 2017

Progression Categories	Counts	Percent	Observed Average	Expected	Infit Mnsq	Outfit Mnsq	Andrich Threshold
Emerging	1675	1.64%	-1.92	-2.13	1.49	1.31	-----
A	4655	4.57%	-1.1	-1.16	1.19	1.15	-3.3
B	7923	7.78%	-0.62	-0.54	0.93	0.93	-1.9
C	14444	14.18%	-0.16	-0.11	0.81	0.84	-1.45
D	16857	16.55%	0.2	0.22	0.87	0.89	-0.63
E	22018	21.61%	0.57	0.52	0.92	0.93	-0.44
F	8188	8.04%	0.91	0.82	0.9	0.96	1.12
G	11997	11.78%	1.12	1.13	1.09	1.16	***0.05***
H	6089	5.98%	1.54	1.51	1.01	1.07	1.45
I	6606	6.48%	1.94	2.05	1.26	1.29	***1.13***
Beyond	1419	1.39%	3.32	3.34	1.48	1.27	3.98

Table 5
Category thresholds and observed average measure scores for Grip and Manipulation - 2017

Progression Categories	Counts	Percent	Observed Average	Expected	Infit Mnsq	Outfit Mnsq	Andrich Threshold
Emerging	1122	1.10%	-2.16	-2.27	1.34	1.42	-----
A	4872	4.78%	-1.01	-1.08	1.16	1.22	-3.11
B	15964	15.66%	-0.46	-0.36	0.85	0.84	-1.67
C	17887	17.55%	0.12	0.13	0.84	0.79	-0.01
D	24606	24.14%	0.54	0.56	0.84	0.8	0.23
E	33114	32.49%	1.12	1.08	0.91	0.93	0.71
Beyond	4357	4.27%	2.22	2.11	1.01	0.98	3.85

Table 6
Category thresholds and observed average measure scores for Crossing the Midline - 2017

Progression Categories	Counts	Percent	Observed Average	Expected	Infit Mnsq	Outfit Mnsq	Andrich Threshold
Emerging	1430	1.40%	-1.74	-2.09	1.32	1.70	-----
A	5787	5.69%	-0.75	-0.81	1.04	1.06	-2.52
B	25413	24.97%	-0.11	0.02	0.90	0.86	-1.39
C	63041	61.94%	0.73	0.70	0.88	0.91	-0.10
Beyond	6110	6.00%	1.83	1.75	0.98	0.96	4.01

Table 7
Category thresholds and observed average measure scores for Following Directions - 2017

Progression Categories	Counts	Percent	Observed Average	Expected	Infit Mnsq	Outfit Mnsq	Andrich Threshold
Emerging	1061	1.04%	-2.25	-2.51	1.66	1.21	-----
A	2462	2.42%	-1.48	-1.52	1.25	1.12	-3.22
B	6699	6.58%	-1.04	-0.90	0.74	0.83	-2.36
C	5863	5.76%	-0.50	-0.48	0.82	0.82	-0.72
D	10671	10.49%	-0.17	-0.19	0.89	0.87	***-1.10***
E	9675	9.51%	0.09	0.05	0.98	0.95	-0.14
F	6431	6.32%	0.28	0.26	0.96	0.97	***0.39***
G	9607	9.44%	0.45	0.47	0.96	0.97	-0.22
H	16427	16.14%	0.72	0.70	0.95	1.00	-0.13
I	10335	10.16%	1.02	0.97	0.97	1.05	1.12
J	12122	11.91%	1.27	1.30	1.21	1.21	***0.79***
K	8483	8.34%	1.77	1.81	1.16	1.08	1.71
Beyond	1926	1.89%	3.00	3.02	1.34	1.05	3.89

Table 8
Category thresholds and observed average measure scores for Letter Naming - 2017

Progression Categories	Counts	Percent	Observed Average	Expected	Infit Mnsq	Outfit Mnsq	Andrich Threshold
Emerging	3264	3.21%	-1.62	-1.80	2.21	1.68	-----
A	5284	5.19%	-1.08	-1.13	1.55	1.76	-1.99
B	5152	5.06%	-0.86	-0.76	0.89	1.18	-0.91
C	2003	1.97%	-0.42	-0.51	1.15	1.27	***0.31***
D	3312	3.26%	-0.25	-0.32	1.27	1.56	***-0.92***
E	6204	6.10%	-0.09	-0.16	1.23	1.51	-0.88
F	3035	2.98%	0.00	-0.01	1.16	1.42	***0.62***
G	4329	4.26%	0.20	0.15	1.23	1.54	-0.30
H	9050	8.90%	0.38	0.33	1.21	1.40	***-0.51***
I	13105	12.88%	0.50	0.55	1.27	1.10	0.05
J	20330	19.98%	0.85	0.84	1.23	1.25	0.24
K	22211	21.83%	1.25	1.28	1.26	1.12	0.94
Beyond	4455	4.38%	2.08	2.22	1.49	1.08	3.36

Table 9
Category thresholds and observed average measure scores for Hand Dominance - 2017

Progression Categories	Counts	Percent	Observed Average	Expected	Infit Mnsq	Outfit Mnsq	Andrich Threshold
Emerging	633	0.62%	-2.60	-2.67	1.07	0.97	-----
A	2589	2.55%	-1.11	-1.03	0.95	0.97	-3.09
B	41130	40.46%	-0.10	-0.11	1.03	1.06	-2.72
C	10118	9.95%	0.37	0.44	0.84	0.78	2.14
D	40604	39.94%	0.98	0.97	0.97	0.97	***-0.14***
Beyond	6585	6.48%	1.80	1.88	1.17	1.14	3.8

Table 10
Descriptive statistics by progression - 2017

	Engagement	Object Counting	Emotional Literacy	Grip and Manipulation	Crossing Midline	Following Directions	Letter Naming	Hand Dominance
Mean	2.89	4.60	4.88	3.73	2.65	6.80	7.92	3.05
Median	3.00	4.00	5.00	4.00	3.00	7.00	9.00	3.00
SD	1.12	2.23	2.24	1.35	0.74	2.96	3.43	1.10
Median Category	C	D	E	D	C	G	I	C

Table 11
Correlations between placements across the progressions - 2017

	Engagement	Object Counting	Emotional Literacy	Grip and Manipulation	Crossing Midline	Following Directions	Letter Naming
Object Counting	.533						
Emotional Literacy	.550	.585					
Grip and Manipulation	.517	.514	.533				
Crossing Midline	.431	.401	.411	.494			
Following Directions	.570	.593	.634	.553	.437		
Letter Naming	.467	.601	.494	.488	.396	.578	
Hand Dominance	.441	.443	.465	.583	.432	.467	.393

Table 12
Item difficulty and discrimination by subgroup - 2017

Progression (scale points)	Index	Total Sample	Female	Male	African American	Hispanic	White
Engagement (0-5)	Classical Item Difficulty	2.89	3.01	2.78	2.65	2.79	3.05
Emotional Literacy (0-10)		4.88	5.03	4.74	4.45	4.49	5.27
Grip and Manipulation (0-6)		3.73	3.89	3.59	3.54	3.68	3.84
Hand Dominance (0-5)		3.05	3.12	2.99	2.93	3.01	3.14
Crossing Midline (0-4)		2.65	2.71	2.60	2.59	2.60	2.70
Object Counting (0-9)		4.60	4.63	4.57	4.16	4.09	5.00
Letter Naming (0-12)		7.92	8.03	7.82	7.43	7.06	8.47
Following Directions (0-12)		6.80	7.02	6.61	6.22	6.30	7.31
Engagement (0-5)	Rasch Item Difficulty	0.07	-0.03	0.16	0.19	0.00	0.02
Emotional Literacy (0-10)		0.54	0.54	0.54	0.54	0.54	0.52
Grip and Manipulation (0-6)		-0.21	-0.30	-0.12	-0.21	-0.36	-0.12
Hand Dominance (0-5)		-0.56	-0.56	-0.56	-0.61	-0.71	-0.48
Crossing Midline (0-4)		-0.45	-0.52	-0.40	-0.52	-0.54	-0.37
Object Counting (0-9)		0.41	0.47	0.36	0.44	0.48	0.38
Letter Naming (0-12)		0.01	0.05	-0.02	-0.03	0.10	0.01
Following Directions (0-12)		0.18	0.18	0.18	0.18	0.18	0.18

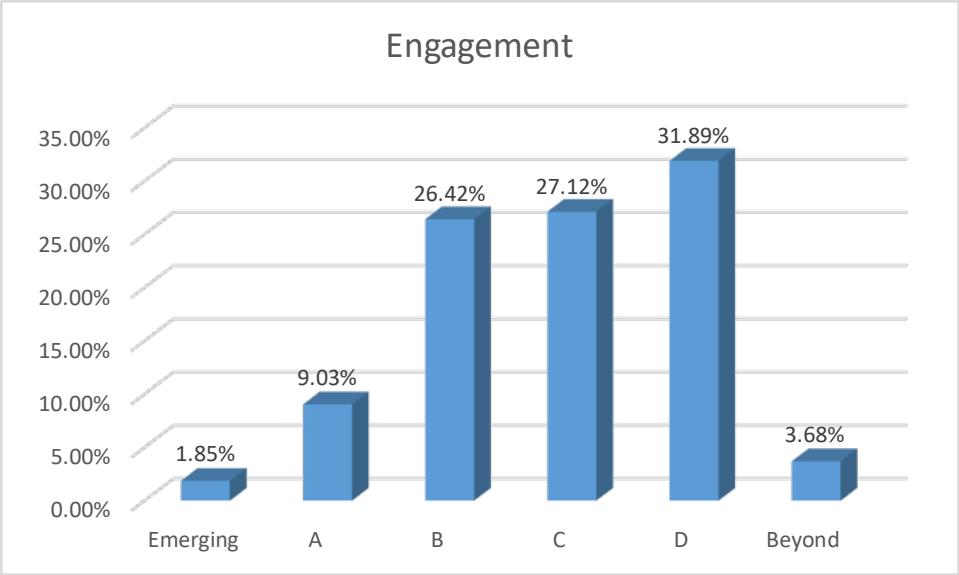

Figure 2. Distribution of progression placements for Engagement – 2017.

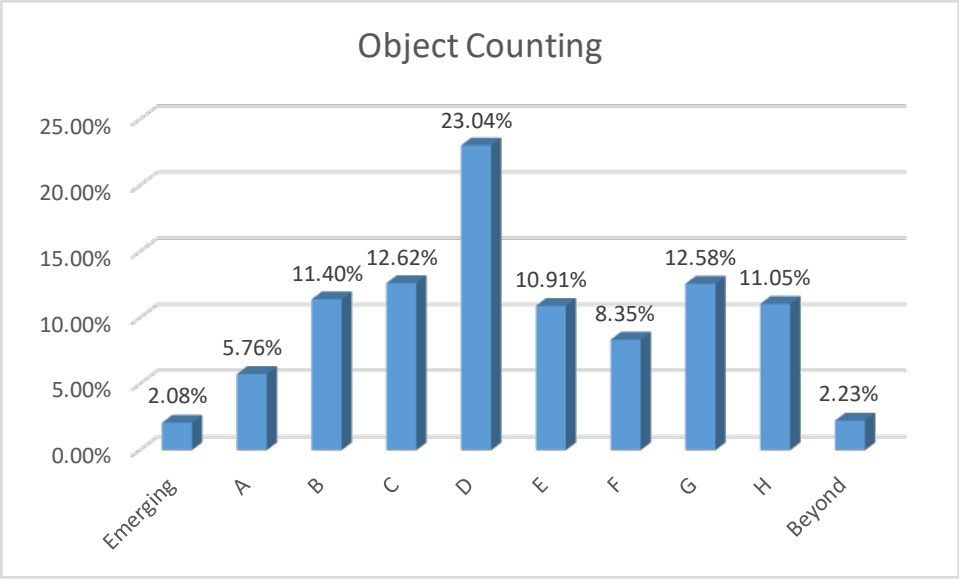

Figure 3. Distribution of progression placements for Object Counting – 2017.

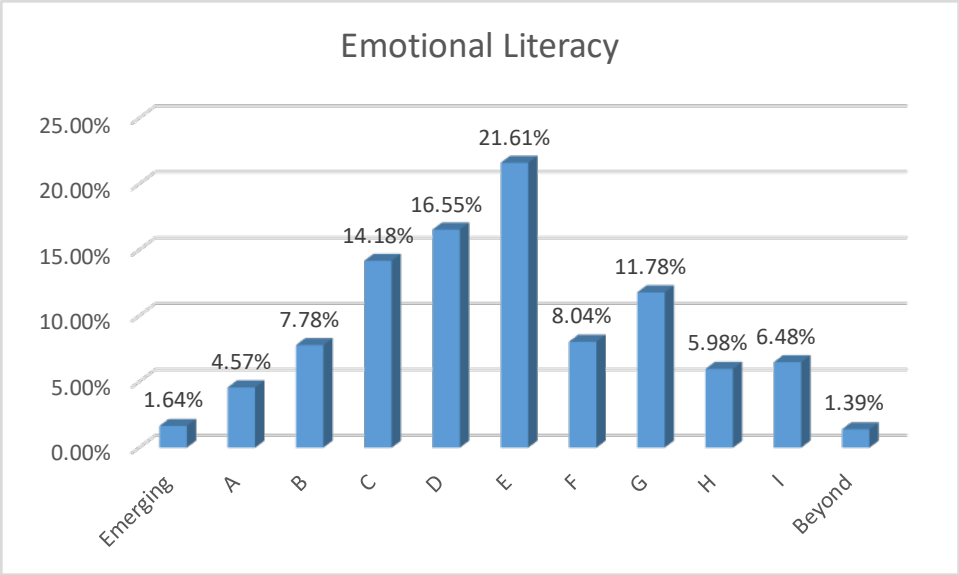

Figure 4. Distribution of progression placements for Emotional Literacy – 2017.

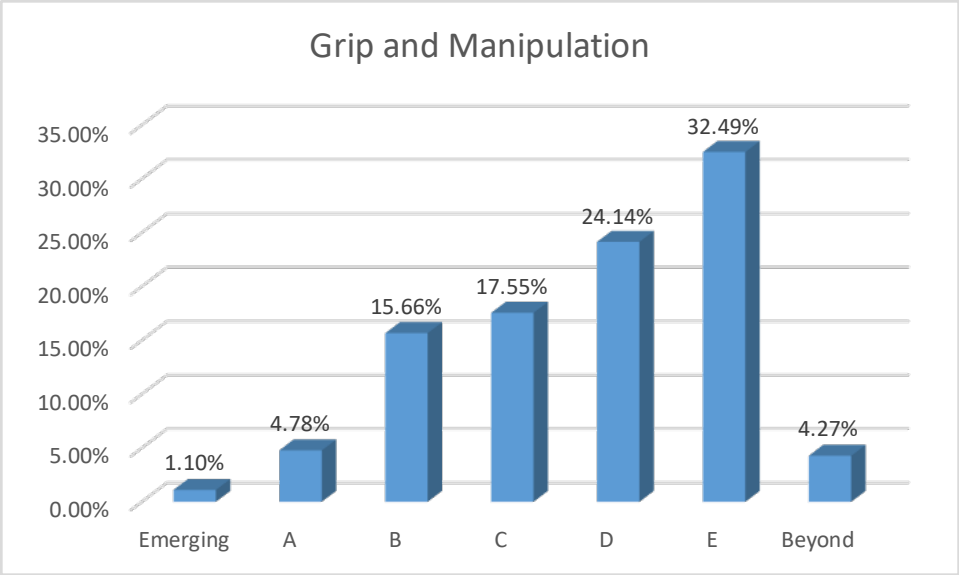

Figure 5. Distribution of progression placements for Grip and Manipulation – 2017.

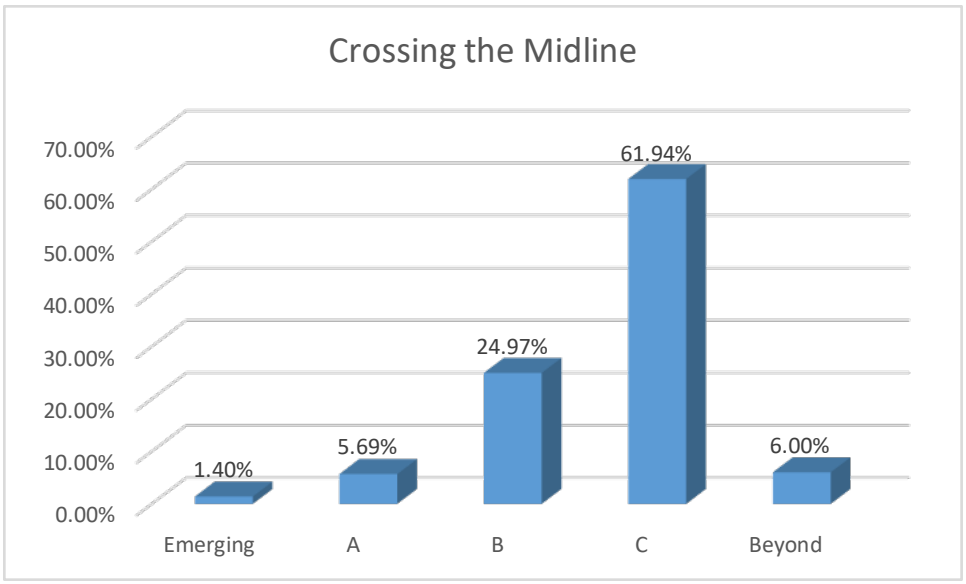

Figure 6. Distribution of progression placements for Crossing the Midline – 2017.

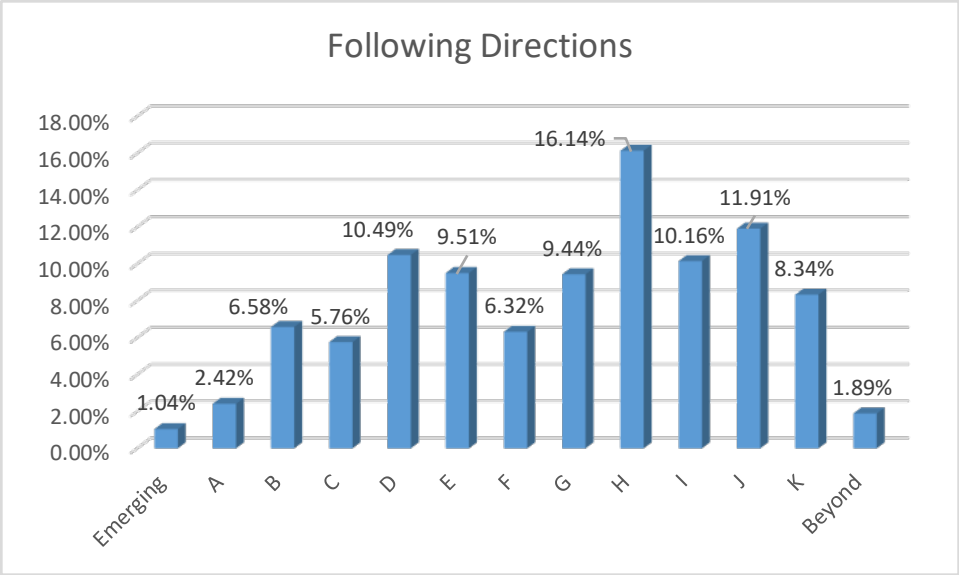

Figure 7. Distribution of progression placements for Following Directions – 2017.

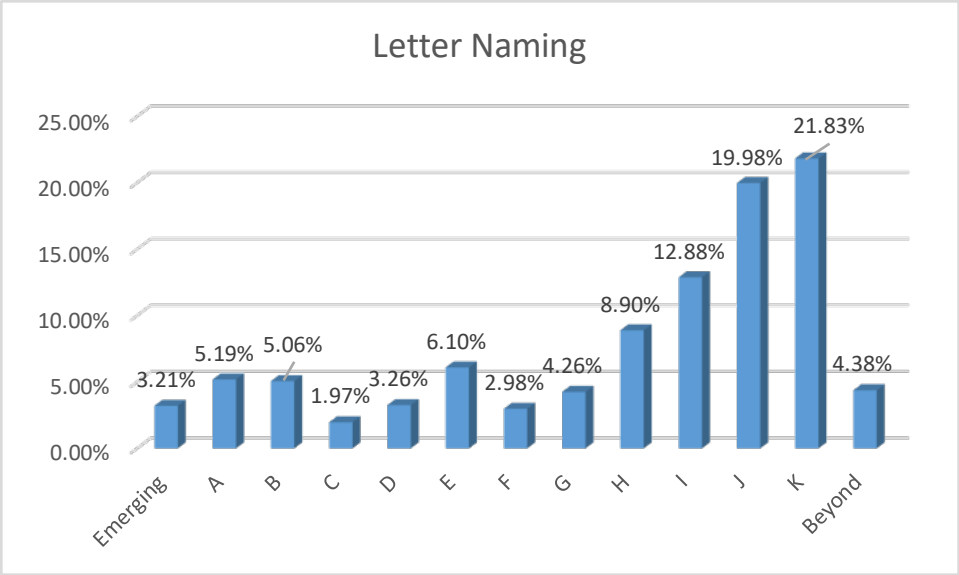

Figure 8. Distribution of progression placements for Letter Naming – 2017.

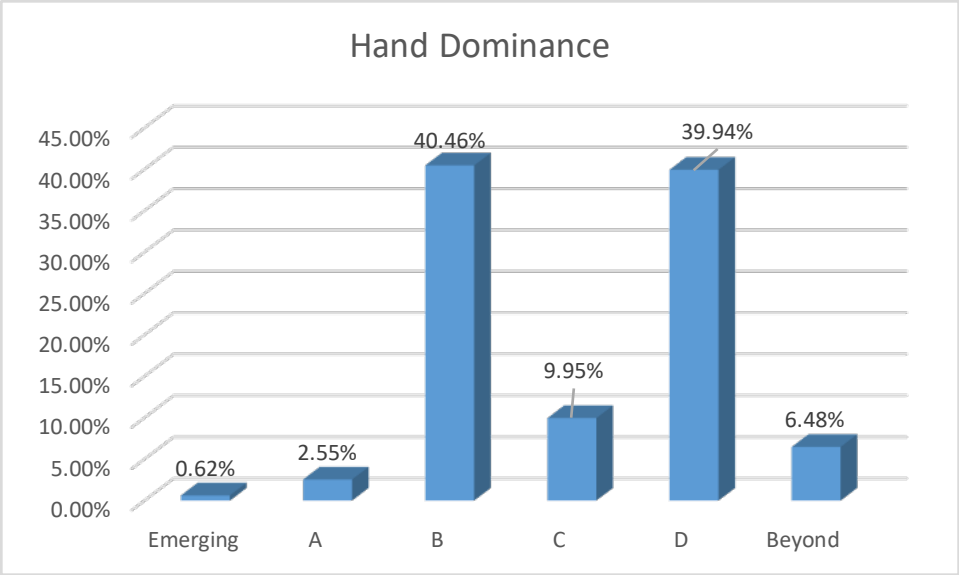

Figure 9. Distribution of progression placements for Hand Dominance – 2017.

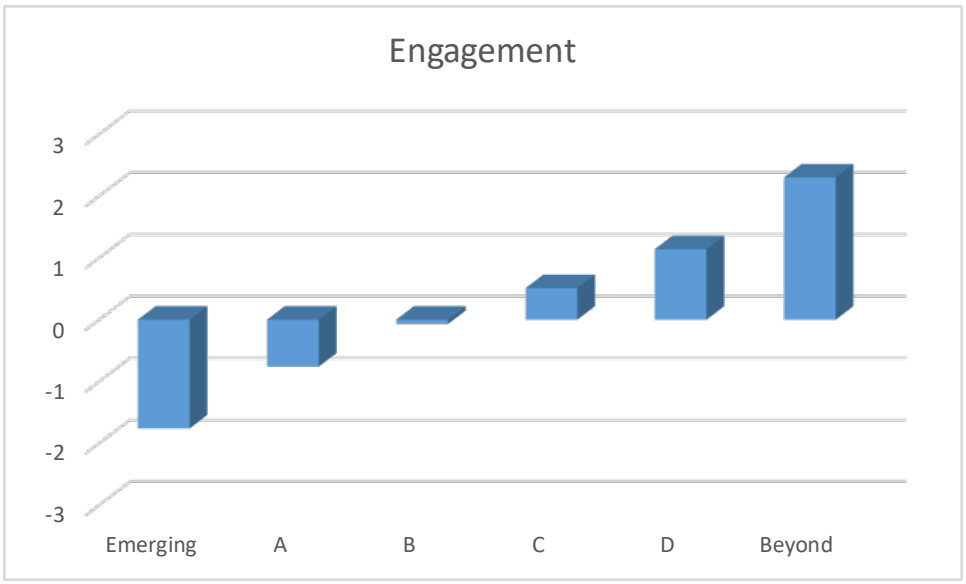

Figure 10. Average Measure Scores by Category for Engagement – 2017.

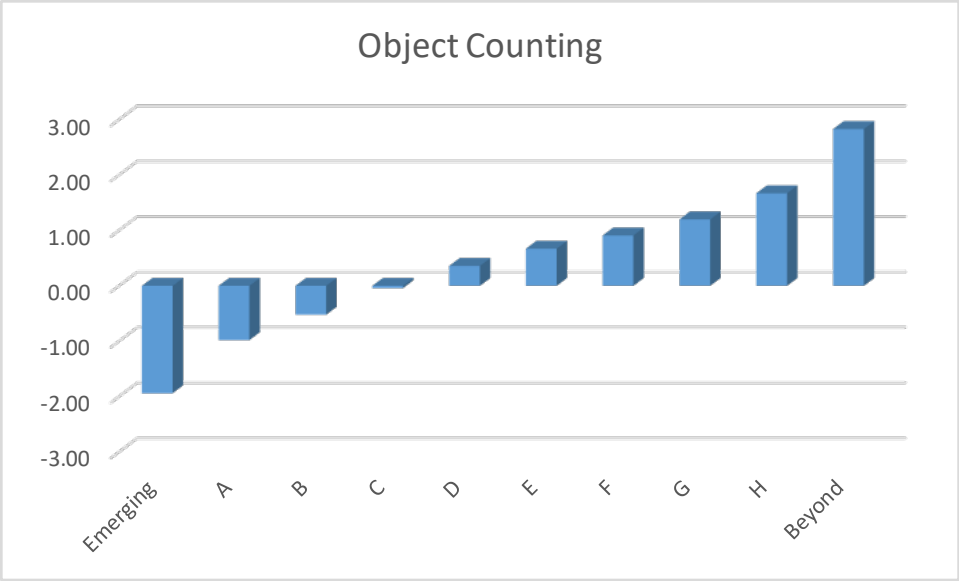

Figure 11. Average Measure Scores by Category for Object Counting – 2017.

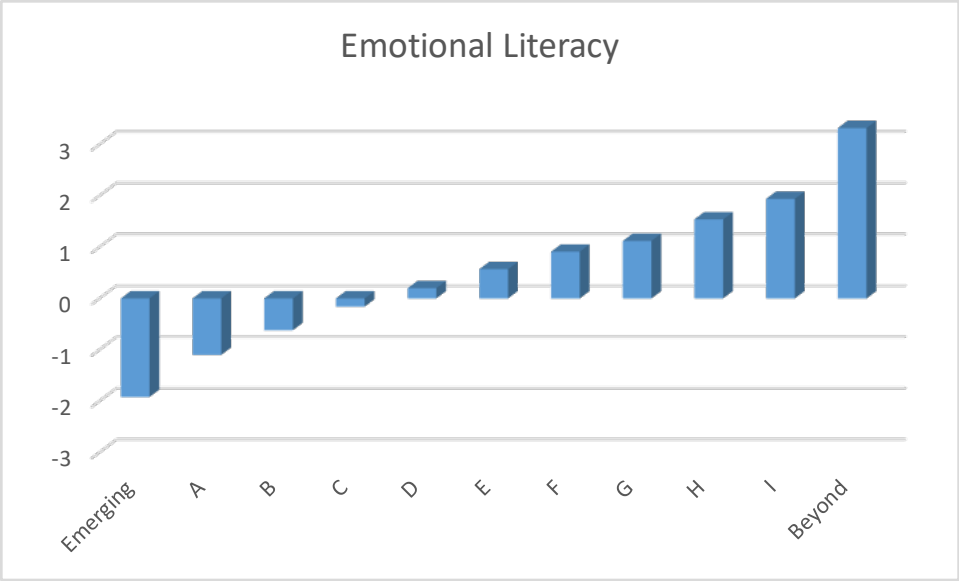

Figure 12. Average Measure Scores by Category for Emotional Literacy – 2017.

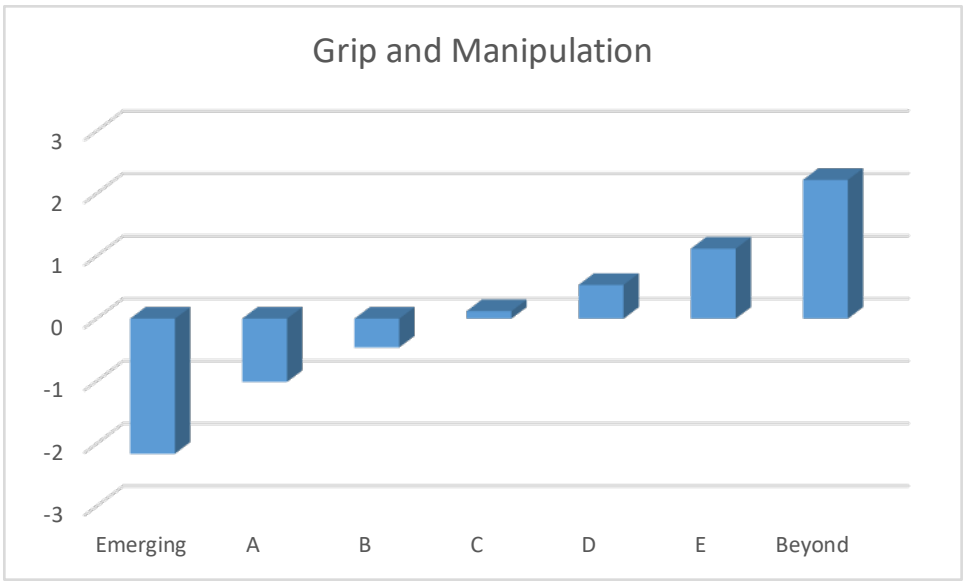

Figure 13. Average Measure Scores by Category for Grip and Manipulation – 2017.

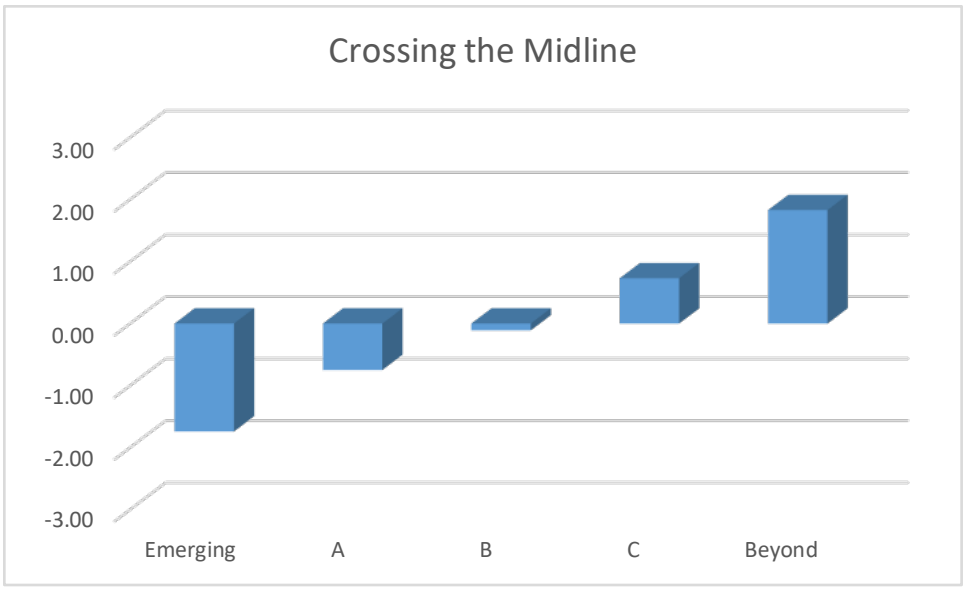

Figure 14. Average Measure Scores by Category for Crossing the Midline – 2017.

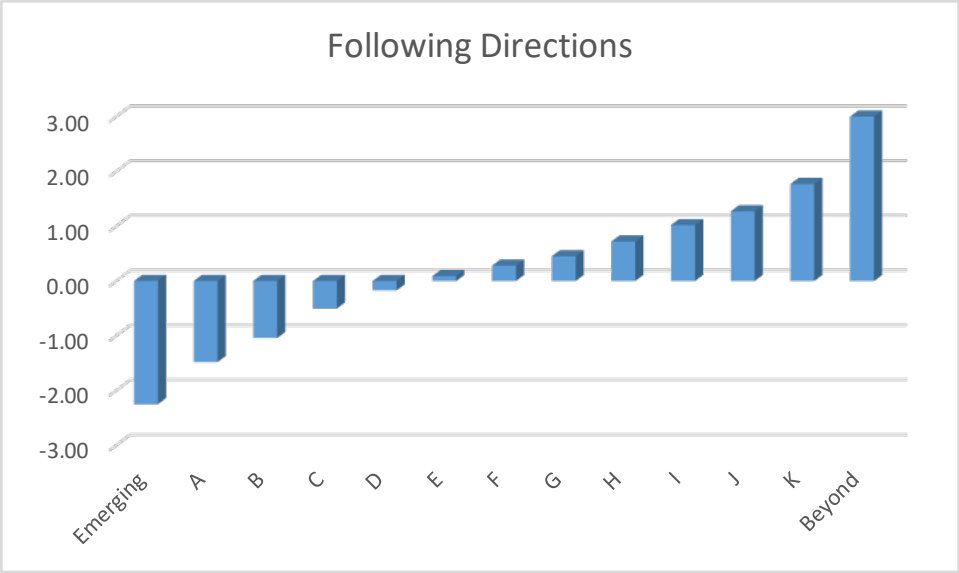

Figure 15. Average Measure Scores by Category for Following Directions – 2017.

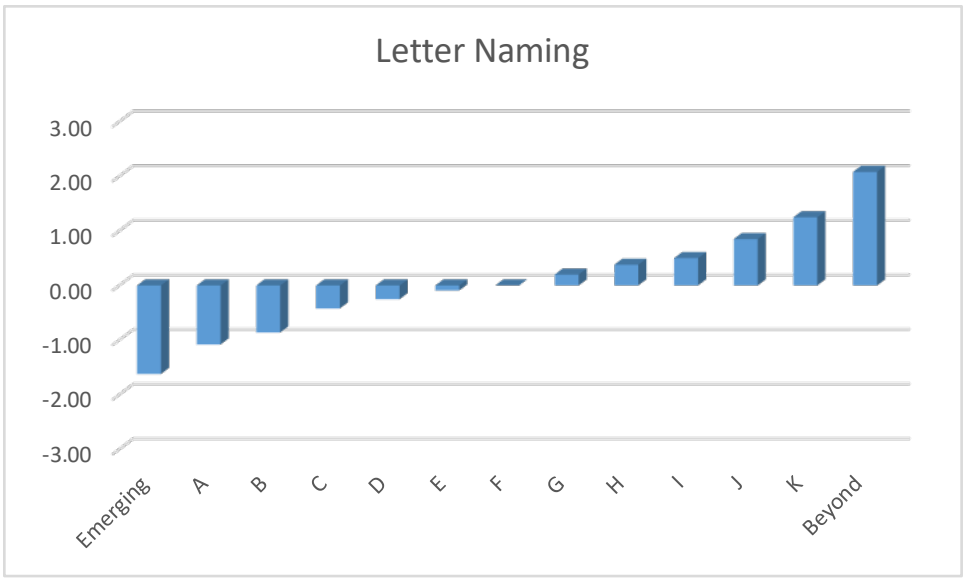

Figure 16. Average Measure Scores by Category for Letter Naming – 2017.

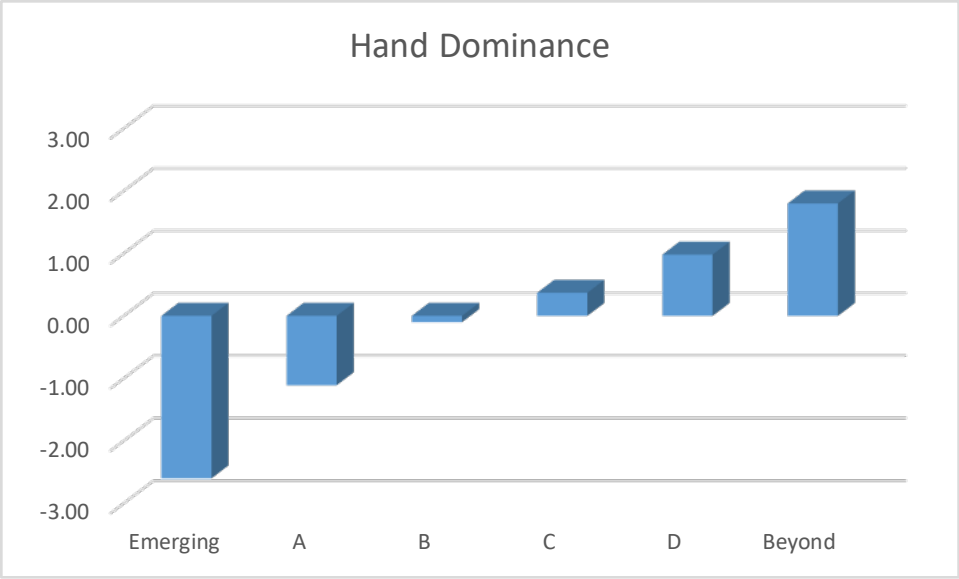

Figure 17. Average Measure Scores by Category for Hand Dominance – 2017.

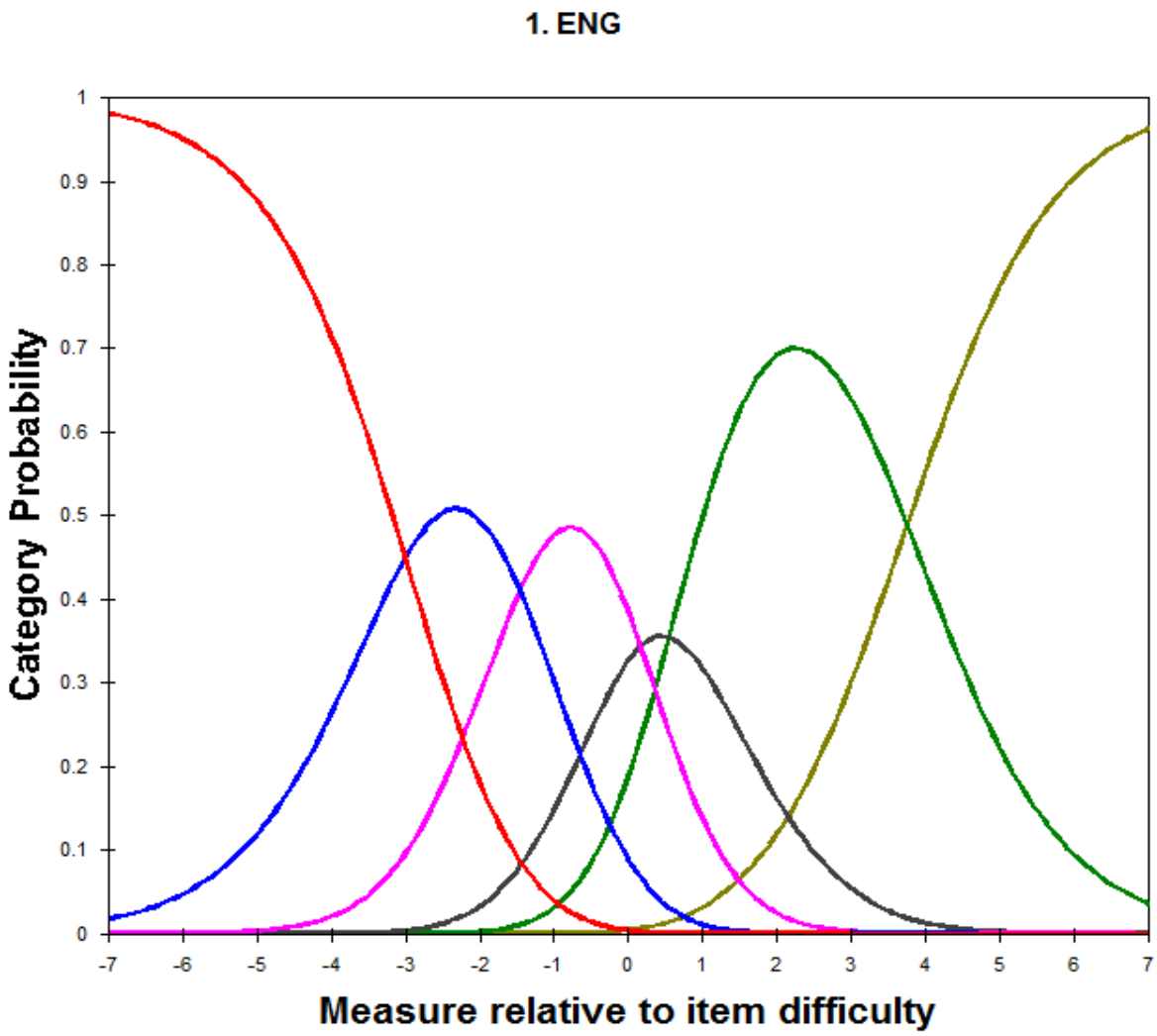

Figure 18. Category Probability Plot for Engagement – 2017.

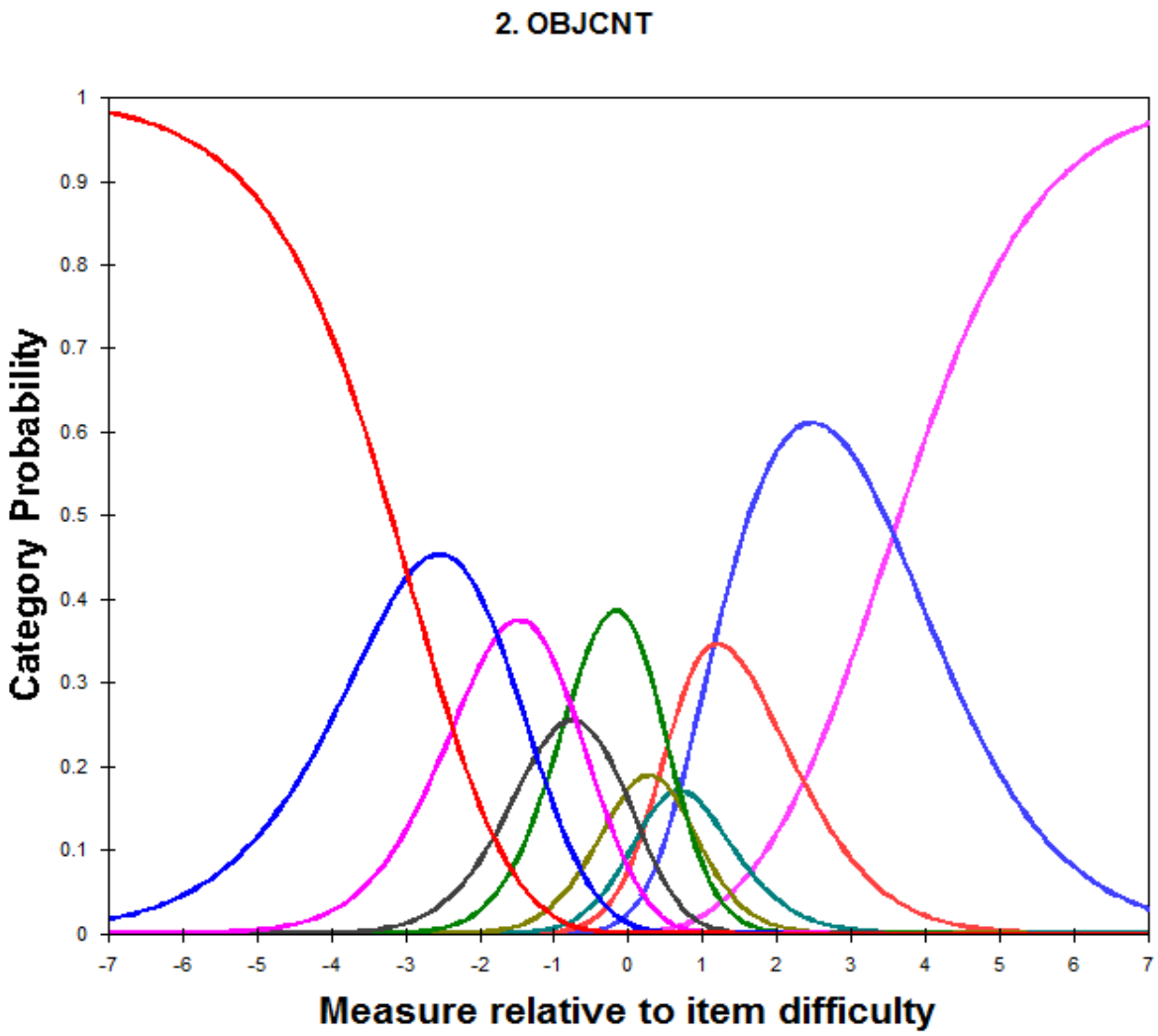

Figure 19. Category Probability Plot for Object Counting – 2017.

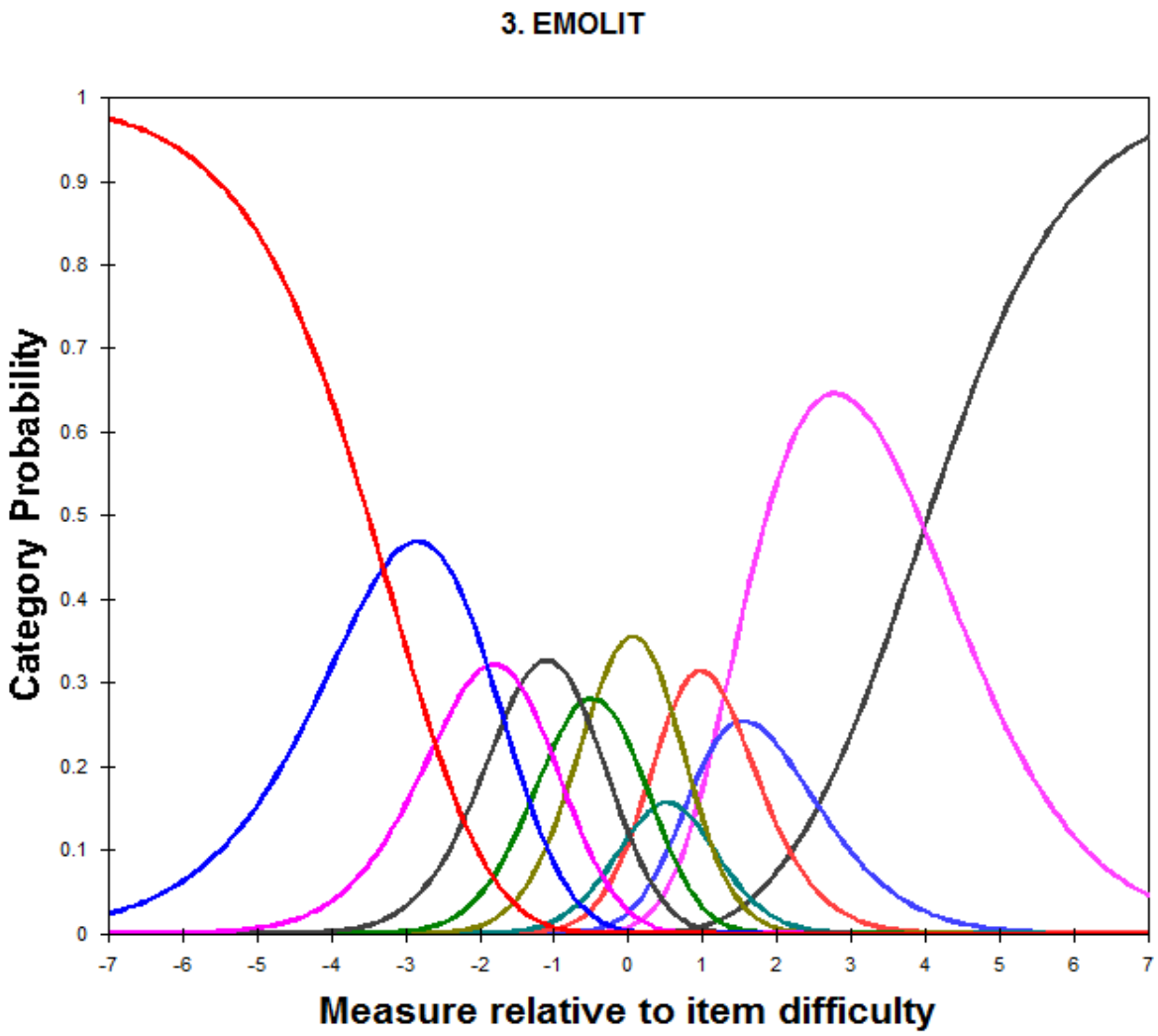

Figure 20. Category Probability Plot for Emotional Literacy – 2017.

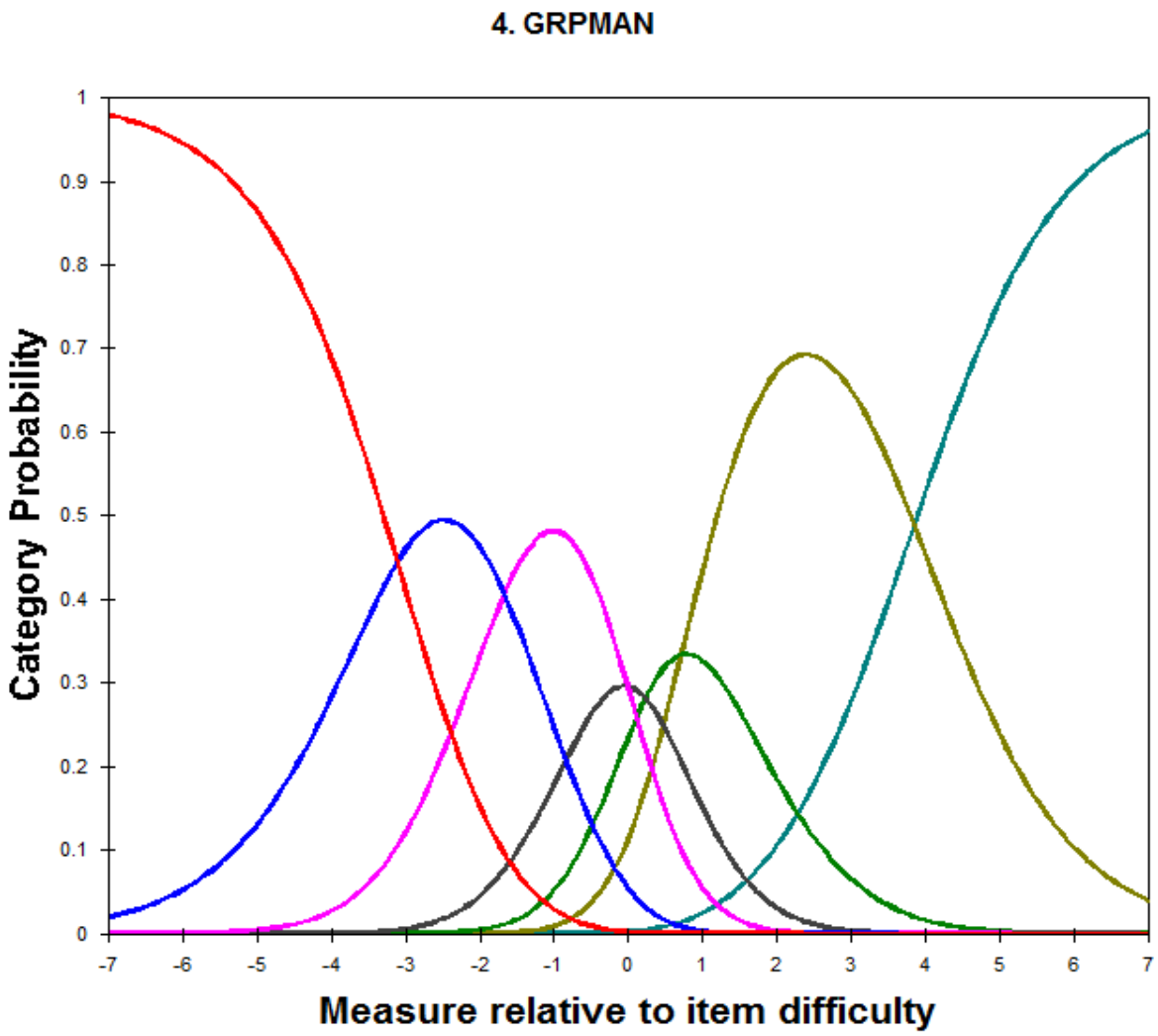

Figure 21. Category Probability Plot for Grip and Manipulation – 2017.

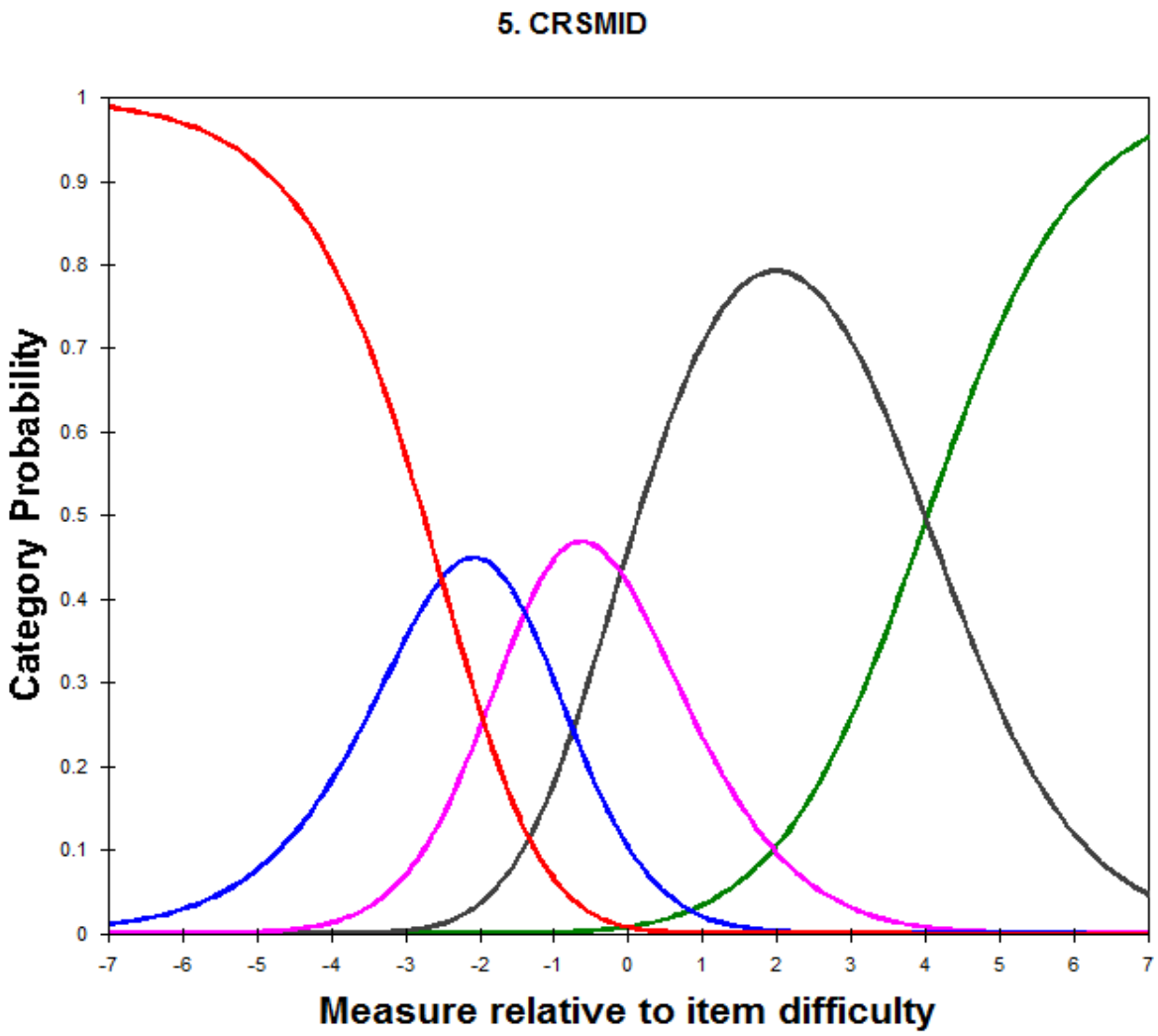

Figure 22. Category Probability Plot for Crossing the Midline – 2017.

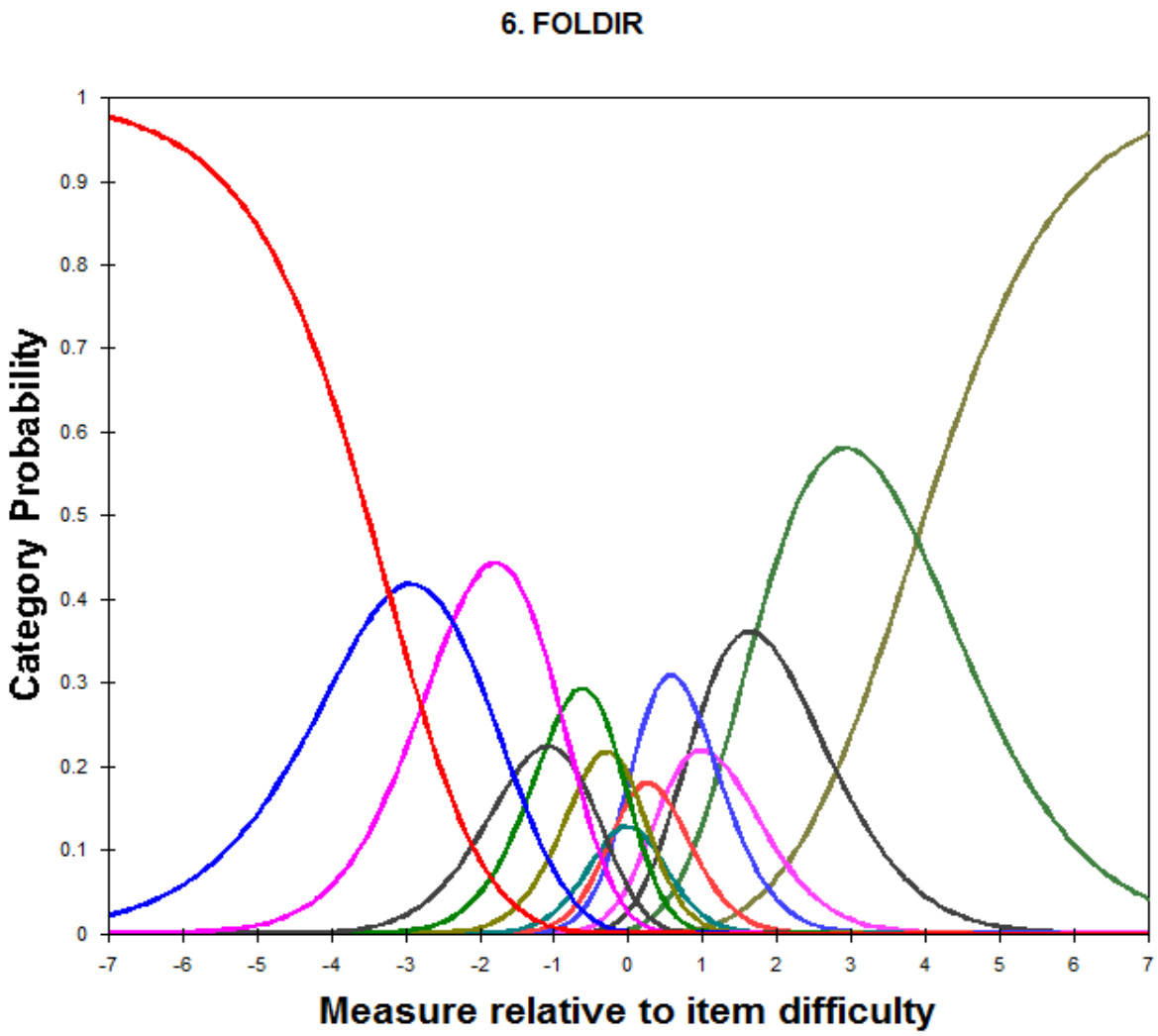

Figure 23. Category Probability Plot for Following Directions – 2017.

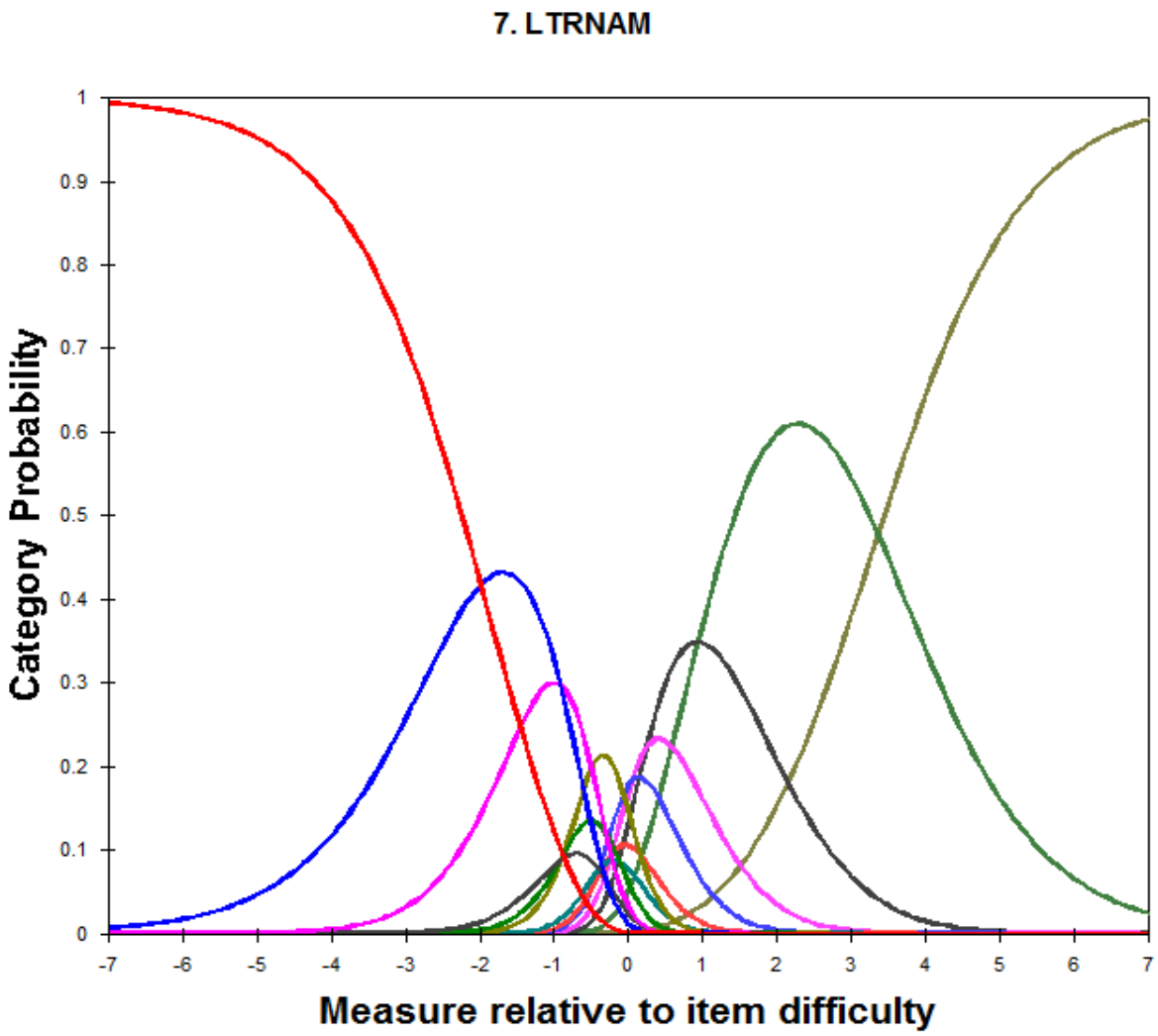

Figure 24. Category Probability Plot for Letter Naming – 2017.

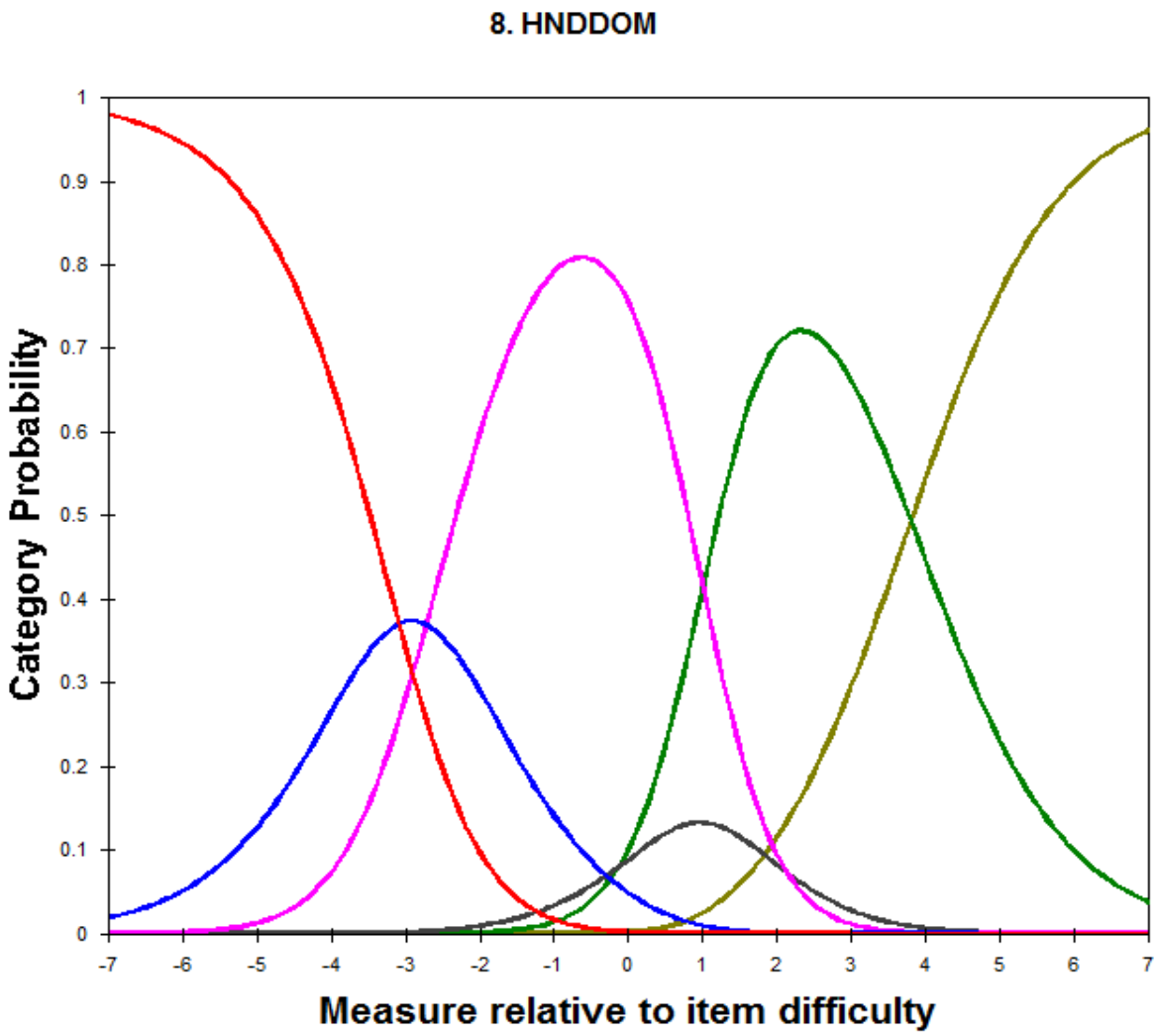

Figure 25. Category Probability Plot for Hand Dominance – 2017.

Summary

It is important to recognize the limitations of these analyses. These analyses focused on only two years of KEA implementation and for many teachers and schools, the process of getting to full implementation is still ongoing. The 2016-17 academic year was only the second year of full state implementation of the KEA and 2017-18 was the third. Many teachers and administrators are still becoming familiar with the KEA progressions and assessment process. The use of all KEA progressions was not required and most teachers used only the required progressions. These analyses are limited to eight of the progressions that were in most common use. Furthermore, these analyses did not examine rater effects, nest the children with their teachers, examine between and within teacher variance, or examine inter-rater reliability. All of these issues will be important topics for future research.

The results of the analyses related to dimensionality and reliability are all very strong and reflect very positively on the use of a total score for psychometric research purposes such as those outlined in this report. The distributions of scores from all of the progressions were moderately correlated with each other (see 2016 and 2017 tables 11). As expected, the lowest correlations were between progressions that would not be expected to be highly related (i.e. Letter Naming and Hand Dominance, 2016 $r = .341$, 2017 $r = .393$). Similarly, the highest correlations were between progressions that would be expected to be related (i.e. Emotional Literacy and Following Directions, 2016 $r = .610$, 2017 $r = .634$). The appropriateness of using a single total score for psychometric diagnostic and research purposes will have to be monitored and evaluated in future studies. There was some indication that multiple factors might be a more appropriate way to treat the data and when all progressions are required there will be more data available to more completely evaluate this issue.

The results related to item or progression difficulty estimates are generally positive, although they suggest a need for a greater range of item difficulties. The results of the examination of rating scale category effectiveness were mixed. There are some very positive results, such as those related to the expected increases in total scores across the categories. There are some mixed results, such as those related to distributional shape and use of the complete scales. There are also some results that indicate cause for concern related to the need for more distinct category probability plots. Future research and continued examination of the psychometric properties of the developmental progressions will be needed to monitor ongoing progress toward full implementation of the KEA assessment as it was intended to be used, and to determine the source of the continuing challenges for teachers.

References

Joint Committee on the Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (AERA/APA/NCME, 2014). *The Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Andrich, D. (1978). Application of a psychometric model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*, 581-594.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Heritage, M. (2013). *Formative assessment in practice: A process of inquiry and action*. Boston: Harvard Education Press.

Linacre, J. M. (2012). Winsteps (Version 3.75.1) [Computer Software]. Chicago, IL: Winsteps.com.

Linquanti, R. (2014). *Supporting Formative Assessment for Deeper Learning: A Primer for Policymakers*. Washington, DC: Council of Chief State School Officers.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Shepard, L. A. (2000). The role of assessment in a learning culture, *Educational Researcher, 29(7),* 4-14.

Zwick, R., Thayer, D. T., & Lewis, C. (1999) An Empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*(*1),* 1-28.