CEMETR-2017-05
DECEMBER-2017

CEME Technical Report

The Center for Educational Measurement and Evaluation

Evidence Supporting the Use of *GOLD*® with Kindergarten Children

Richard G. Lambert

RICHARD LAMBERT
CHUANG WANG
MARK D'AMICO
SERIES EDITORS

A PUBLICATION OF THE CENTER FOR EDUCATIONAL MEASUREMENT AND EVALUATION

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

Evidence Supporting the Use of *GOLD*® with Kindergarten Children

Richard G. Lambert, Ph.D.

Center for Educational Measurement and Evaluation

UNC Charlotte

December, 2017

Teaching Strategies GOLD® Assessment System (*GOLD®*) is a formative assessment system that has been designed and extensively validated for use with young children ages birth to kindergarten. For a thorough review of the process of developing the measure and the existing research evidence to support the use of the measure, see existing research articles and the 3rd edition of the *GOLD®* Technical Manual (Lambert, Kim, & Burts, 2013; Lambert, Kim, & Burts, 2014; Lambert, Kim, & Burts, 2015). This report focuses on establishing reliability and validity evidence for the Birth to Third Grade (B-3) version of the assessment system when used with kindergarten children.

The *GOLD®* measure yields information that is rooted in the ongoing work of teachers as they develop and collect evidences that are used to identify the best fits for each child across a series of developmental progressions. Teachers collect ongoing portfolios of evidences throughout the academic year, reflect upon and analyze those evidences, make preliminary ratings on an ongoing basis, and finalize ratings at specified points during the 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. *GOLD®* helps 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 provided by *GOLD*® using a nationally representative sample of kindergarten children.

Background Information on the Development of *GOLD*®

GOLD® (Heroman et al., 2010) measures the progress of children ages birth through third grade in the major developmental and content areas. The objectives help teachers organize their documentations as they regularly gather information through observations, conversations with children and families, samples of children's work, photos, video clips, recordings, etc. Teachers summarize child assessment information at three checkpoint periods during the year (i.e., fall, winter, and spring). The information is intended to be used to assist teachers in planning appropriate experiences, individualizing instruction, and monitoring and communicating child progress to families and other stakeholders. *GOLD*® is intended for use with typically developing children, children with disabilities, children who demonstrate competencies beyond typical developmental expectations, and dual language learners.

The development of *GOLD*® occurred over several years and incorporated feedback from teachers, administrators, consultants, and Teaching Strategies, LLC professional-development and research personnel. Pilot studies with diverse populations were conducted, and a draft of the measure was sent to leading authorities in the field for content review. Major revisions were made based on results of the content validation and pilot studies. Final assessment items were selected on the basis of feedback received during the development process; state early learning standards and the *Head Start Child Development and Early Learning Framework* (U.S. Department of Health & Human Services, 2010); and current research and professional literature including literature that identifies which knowledge, skills, and abilities are most predictive of school success. This process resulted in a total of 38 objectives with 23 of them in the areas of social-emotional, physical, language,

cognitive, literacy, and mathematics. *GOLD*® also includes objectives in other areas (i.e., science and technology, social studies, the arts, and English language acquisition).

Objectives in the social–emotional domain involve understanding, regulating, and expressing emotions; building relationships with others; and interacting appropriately in situations. The physical domain objectives include gross-motor development (traveling, balancing, and gross-motor manipulative skills) and fine-motor strength and coordination. The language objectives include understanding and using language to communicate or express thoughts and needs. Objectives in the cognitive domain include approaches to learning (e.g., attention, curiosity, initiative, flexibility, problem solving); memory; classification skills; and the use of symbols to represent objects, events, or persons not present. The literacy objectives incorporate phonological awareness; alphabet, print, and book knowledge; comprehension; and emergent writing skills. The mathematics objectives focus on number concepts and operations, spatial relationships and shapes, measurement and comparison, and pattern knowledge.

The *GOLD*® measure has been expended to include more rating scale items and additional rating scale categories in order to incorporate developmental expectations for children up to third grade. The 23 *GOLD*® objectives included in the current studies are now operationalized into 60 rating scale items: social–emotional (9 items), physical (5 items), language (8 items), cognitive (10 items), literacy (16 items) and mathematics (12 items). Teachers rate children's skills, knowledge, and abilities along rating scales that range from 10 to 19 points and outline progressions of development and learning. These progressions range from "Not Yet" (Level 0) to "Exceeds Third Grade Expectations" (Levels 9 to 19 depending upon the progression). Each progression includes indicator levels with varied examples from everyday situations that give teachers guidance of what evidence may look like. There are also "In-between" levels and do not include examples. They allow for additional steps in the progression as the child demonstrates that skills are emerging in a particular

area but are not fully established. Overlapping, color-coded bands indicate the typical age and/or grade-level (i.e., kindergarten) ranges for each item measured.

Background Information on the Validation of *GOLD*®

The psychometric properties of *GOLD*® have previously been explored for its use with children representing different ethnic, racial, language, functional status, and age groups. These initial studies suggest that *GOLD*® is a psychometrically promising instrument which has utility for children representing diverse populations. High internal consistency reliability (*a*s = .95 - .99) and moderately high Rasch reliability statistics (person separation = 9.42, item separation = 19.20, person reliability = .99, item reliability = 1.00) were found using a sample (n=290) of infants through children two years of age (Kim & Smith, 2010).

Lambert, Kim, & Burts (2012) explored the (a) factorial structure of the *GOLD*®, (b) indexes of reliability, and (c) inter-rater reliability. Findings suggested that the *GOLD*® measures six separate domains as intended. Inter-rater reliability between a master trainer and teachers was high. Reliability coefficients for all three checkpoints were also high. Results of longitudinal invariance CFA indicated the constructs were equivalent across time implying that the interpretations of changes in children's development and learning obtained from the measure are valid.

Another study looked at the validity of *GOLD*® for assessing children with disabilities and those for whom English is not their first language. Assessment information was collected on three-, four-, and five -year-old children at the fall (n=79,324), winter (n=132,693), and spring (n=50,558) checkpoints. Differential Item Functioning (DIF) analysis indicated that in general, teachers' ratings were similar for children of similar abilities, regardless of their subgroup membership. The majority of items in the *GOLD*® displayed little or no Differential Item Functioning (DIF) with the exception of one item, "uses conventional grammar" (Kim, Lambert, & Burts, 2013).

Associations of teacher ratings with child demographics (e.g., age, gender, disability status, English language status) and classroom composition characteristics (e.g., class mean age and percentage ELLs, children with disabilities, and males) were examined with a sample of 21,592 children ages 12 months through 59 months. Using three-level growth curve modeling, findings indicated that teachers' *GOLD*® ratings were associated in anticipated directions for both child and classroom characteristics. Children with disabilities began the year behind their typically developing peers and grew more slowly throughout the year. Girls demonstrated advantages in some areas over boys. ELLs were rated lower at the beginning of the year but exhibited somewhat faster growth rates than native English-speakers. Differences in rater effects (i.e., how teachers used the *GOLD*® to rate the children in their classrooms) ranged from 16% to 25%, which is considerably lower than reported in some studies (Lambert, Kim, & Burts, 2013).

The dimensionality, rating scale effectiveness, hierarchy of item difficulties, and the relationship of *GOLD*® developmental scale scores to child age have also been examined. Data from a norm sample (n=10,963) of children ages birth to 71 months were analyzed using the Rasch Rating Scale Model to develop interval level scale scores that could be used to track children's development and learning across the intended age range. Support was found for the unidimensionality of each domain (i.e., items in each scale measure one and only one underlying latent construct). Results further indicated that teachers can make valid ratings of the developmental progress of children across the measured age range. Correlations were moderately high between each of the scale scores and child age in months with correlation coefficients ranging from .67 to .73. The rating structure functioned effectively with the exceptions that ratings at the lowest and highest ends of the scale were somewhat less reliable and in-between ratings were less distinct. Overall, items formed theoretically expected hierarchies such that items which were less difficult for children were rated by teachers as less difficult (Kim, Lambert, & Burts, 2014).

A preliminary study of *GOLD*® with a subsample of infants through children two years of age (Kim & Smith, 2010) indicated high internal consistency reliability (*a*s = .95 - .99) and moderately high Rasch reliability statistics (person separation = 9.42, item separation = 19.20, person reliability = .99, item reliability = 1.00). Concurrent validity using a modified version of the *GOLD*® (i.e., *WaKIDS*) with kindergarten children (n=333) was explored by researchers in Washington state. Moderate correlations (*r* =.50 - .64) with a battery of established norm-referenced achievement instruments were found for the Language, Literacy, and Mathematics areas (Soderberg, Stull, Cummings, Nolen, McCutchen, & Joseph, 2013).

The first version of the technical manual for the *Teaching Strategies GOLD*™ Assessment System (Lambert, Kim, Taylor, & McGee, 2010) presented initial reporting of reliability and validity evidence based on the information the measure provides to teachers of young children. The manual contained evidence concerning the dimensions measured by the assessment system and their interrelationships. The results outlined the measurement model used to create scale scores for each dimension. The report also contained a variety of strong statistical evidences concerning the fit of the data provided by the assessment system to the measurement model. Strong reliability evidence was presented from both classical and modern indexes of internal consistency, along with the results of a study of inter-rater reliability. Norm tables for each scale score were provided based on three month age bands spanning ages 6 to 71 months.

At the time the initial manual was produced, the assessment system was relatively new and many of the teachers had been using the system for only one year. Since the last report, many more states and programs have adopted the assessment system, much more training has taken place, and more research has been conducted on the system. Since *GOLD*® was released in the fall of 2010, the number of teachers using the tool has grown dramatically, with over 2.5 million child portfolios have been gathered. All teachers have access to free training through the online courses, as well as

Inter-rater reliability checks. In addition to the free training, thousands of teachers are trained each year, using face-to-face training, to ensure their knowledge of how to use the tool. *GOLD*® is widely used in all states for Pre-k assessment and in many states for Kindergarten entry assessment.

Given the widespread use of *GOLD*®, greater availability of teacher training, and much more sophisticated and experienced use of the system, a second technical manual was produced to provide an updated set of evidences based on an up-to-date nationally representative norm sample that reflected how *GOLD*® was being used. The revised manual (Lambert, Kim, & Burts, 2013) provided updated reliability and validity evidence based on both classical and Item Response Theory based measurement models. Norm tables were provided that covered children aged birth through 71 months. For each age band, expected scores for the fall, winter, and spring assessments, age specific standard errors of measurement, and expected growth from fall to spring were provided for both standard scores and raw scores.

The Purpose of the *GOLD*®

GOLD® has been designed and validated to be used as a formative, developmental, authentic, and criterion referenced classroom measurement tool for teachers. By extension, it is not a 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 assessments 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 both the *GOLD*® measure and the information it provides. The following section will attempt to help define the purpose of *GOLD*® and how it can be a helpful resource for teachers and those who support teachers and children. We will contrast optimal and ineffective uses

and outline the most meaningful and appropriate applications of the measure.

***GOLD*® is a Formative Assessment**

Formative assessment focuses on the learning process and is used to support learning while learning is taking place. *GOLD*® has been designed for formative purposes. 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 help students improve their achievement of intended instructional outcomes…" (CCSS, 2006; AERA/APA/NCME, 2014). *GOLD*® 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. It can also help teachers understand the current status of the growth, learning, and development of the children in their classrooms at every point in the academic year. This information can help teachers understand the interests of the children, plan classroom activities, select and rotate classroom materials, and individualize and differentiate instruction.

The *GOLD*® measure 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. This process can help facilitate rich conversations about the child's development and their family and cultural context. Teachers can even solicit evidences of child progress and development from parents and other caregivers. This process can help teachers partner with parents to support the growth and development of each child.

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.

This information can also be useful to those who support the professional development of teachers. It can provide an enhanced picture of how a teacher experiences and is aware of classroom processes, observes children in the classroom, and collects evidences of child progress and development. Formative assessment data can provide process information regarding how teachers analyze evidences of child progress, make placements on developmental progressions, and use that information to support child learning, growth, and development. This same data can be very useful as mentors and other support personnel help teachers plan individual, small group, and whole group instructional activities.

Mentors, coaches, and technical assistance providers can also use formative assessment data as a catalyst for rich conversations with teachers that can help them reflect about instructional practice and set professional development goals. This process can help teachers increase their observational skills and become much more aware of how each child learns and functions in the classroom. In this way, mentors can provide data driven support to teachers and thereby model for teachers the process of using data to individualize support for children.

Formative assessments are not developmental screeners. *GOLD*® has not been designed to provide cut scores that indicate the need for further testing or diagnostic processes. Similarly, it has not been designed to lead to specific decisions using a cut score that result in high correct classification or high false positive rates. Formative assessment information alone is not appropriate for making high stakes placements or diagnostic classifications of children and no such decisions should be based on single sources of information. However, the information provided by high quality use of well-developed and validated formative assessments can make valuable contributions to multiple source, multifaceted, multidimensional and multidisciplinary professional discussions of the needs of individual 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.

***GOLD*® is a Developmental Assessment**

GOLD® has been designed to be developmental, meaning that it includes progressions of growth, development and learning that describe a sequence of stages that children are generally expected to demonstrate. Each progression includes descriptive anchors that illustrate behaviors, work samples, and other evidences that can be observed in the classroom. It is designed to help

teachers learn about and understand the whole child. It can help provide information across multiple domains of development and is sensitive to child growth and development over time.

GOLD® has been aligned to the Head Start Early Learning Outcomes Framework, Common Core Standards and many state early learning standards. The most commonly used standards throughout the U.S. outline specific skills and abilities that children are expected to obtain by the end of particular grade or age levels. *GOLD*® expresses these standards not just in terms of the destination or end point, but in terms of the journey or learning process. For each standard, a set of instructional objectives has been outlined and in turn each instructional objective is associated with least one developmental progression. These progressions describe in detail the steps children are expected to go through on their way to mastery of new skills and abilities. In this way, the measure helps teachers reflect about and understand more fully the learning standards, curricular goals, and instructional objectives in terms of evidences that can be observed within everyday interactions and activities in the classroom.

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.

GOLD® is not a benchmark assessment. There are not correct and incorrect answers to a set of questions or test items, and it is not designed to indicate which children are or are not on track to achieve specific summative assessment scores at a specific fixed future assessment date. Rather, *GOLD*® helps teachers understand the developmental status of children where ever the children are developmentally. Each developmental progression includes a wide range of behavioral and observational anchors that extend above and below each age expectation level so as to include opportunities to document child growth and development for all children within the intended age ranges.

***GOLD*® is 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 *GOLD*® provides is 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, *GOLD*® 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. Furthermore, direct assessment focus on specific measureable constructs and behaviors. *GOLD*® can help teachers examine the whole child across a variety of developmental domains.

In contrast to direct assessments, *GOLD*® as an authentic assessment relies on teacher skill and professional judgement as applied to the analysis of a rich portfolio of evidences and experiences with children across a wide variety of classroom situations and circumstances. Therefore, there is no formal administration protocol for *GOLD*®. Rather, as with all authentic assessments, administration is an ongoing process through which teachers observe children in their natural classroom environment, and collect work samples, artifacts, evidences, and anecdotal records that describe and illustrate child learning and developmental progress.

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.

***GOLD*® is a Criterion Referenced Assessment**

Criterion referenced measures assess progress and learning relative to a fixed set of standards. They are not designed specifically to spread out children relative to each other along a continuum of achievement at particular point in time. Rather, they are designed to place children along a continuum of growth and development. The information provided by *GOLD*® is most useful for identifying where a given child is functioning relative to their own past developmental trajectory and relative to standards for children of a given age range. These standards are called Widely Held Expectations. They are designed to be used in combination with color bands that represent specific years of age. Teachers can use the Widely Held Expectations for specific age bands to understand what behaviors and skills children of a certain age can generally be expected to demonstrate in the classroom.

Criterion referenced assessment tools are not norm referenced tools. *GOLD*® has not been designed and validated primarily to indicate where a specific child is functioning relative to all other

children of similar age. For example, percentile scores are not the focus of the information provided to teachers. Rather, the focus of *GOLD*® reporting is on specific skills, abilities, and developmental steps. Each developmental progression includes Widely Held Expectations which function as interpretation guidelines. These criteria help teachers understand how each child is growing and developing relative to what is expected for a given age. These expectations are not rooted in quantitative norms that describe how children of a given age have scored on the measure at a fixed point in time. Rather, they are based on developmental theory, expert recommendations, and child development research. The normative information that is available is designed to provide teachers with an additional interpretation resource, and can provide general information for teachers who are interested in a broad and comprehensive picture of how a child is growing and developing relative to both criteria based on Widely Held Expectations and the developmental progress of other children of similar characteristics.

In conclusion, *GOLD*® provides assessment information in support of the ongoing learning process. When properly implemented, it can help teachers gather and analyze evidences of child progress in the natural classroom context while child learning, growth, and development is taking place. It is rooted in evidences that are collected from child work samples and behaviors. These evidences emerge from daily classroom routines and activities, and reflect child behaviors that take place in the natural context of daily instruction in the classroom. It can provide information that is useful for instructional planning and communicating about child progress. It also offers interpretation guidelines that can help teachers understand how a child is developing and progressing relative to research-based, age-specific indicators of developmental progress.

The Current Study

GOLD®, along with a variety of other commercially available and locally developed formative assessment systems, is being widely used as both a kindergarten entry assessment and as a resource to track the growth and development of kindergarten children across the academic year. Teachers throughout the United States are being asked to implement many assessment systems and assessment related tasks. They are attempting to do so while also facing challenges related to limited instructional resources, many demands on their limited time, and increasingly challenging classrooms of children with diverse needs. All of this comes at a time when they are often unprepared for the linguistic and cultural diversity that is the reality of the American classroom of young children. Therefore, they need high quality training and preparation as well as high quality assessment resources to help them document and understand the developmental status, strengths, needs, and growth patterns of the children in their classrooms. The purpose of this study was to examine the psychometric properties and overall quality of *GOLD*® for use with kindergarten children.

National Sample

The data from kindergarten children, in order to be eligible for this study, had to include valid assessments for fall, winter, and spring checkpoints to be eligible for this study. From the total population of kindergarten children assessed using *GOLD*®, a sample was selected that met this criterion. The 2016 Census Bureau national estimates for the proportion of children ages birth to 6 years of age in each ethnicity / race group were compared to the characteristics of the children in the sample. Teachers are required to enter into the *GOLD*® online system information regarding each child's race and ethnicity. The questions about each child are similar to those used by the U.S. Census Bureau. Given that Hispanic identity is an ethnicity, not a racial grouping, and given the importance of representing children of Hispanic ethnicity in the norm sample, the race and ethnicity variables were combined into the following seven ethnic subgroups: 1.) White, not Hispanic; 2.)

African-American; not Hispanic; 3.) Native American or Hawaiian / Pacific Islander, not Hispanic; 4.) Asian, not Hispanic; 5.) multiracial / other, not Hispanic; and 6.) Hispanic.

The sample included a total of 21,258 kindergarten children. These children received educational services in centers or schools that were located in five states: Colorado, Georgia, Maryland, New Jersey, and Tennessee. As shown in Table 1, the norm sample was very evenly balanced by gender (boys=49.1%, girls=50.9%). Children with an IEP comprised 1.3% of the sample. A total of 6.7% of the norm sample qualified for the National School Lunch Program (free or reduced price lunch). The primary language spoken in the home was distributed as follows: English (76.6%), Spanish (19.9%), and other languages (3.5%). The race / ethnicity of the children in the sample was as follows, show here with the national census figures in parentheses: a.) White – 49.6% (50.1%), b.) African American – 8.2% (13.9%), c.) Native American / Pacific Islander – 0.7% (1.2%), d.) Asian – 2.6% (4.8%), e.) Multiracial / other – 16.5% (4.1%), and f.) Hispanic – 22.4% (25.9%). These values indicate that the sample was approximately representative for White and Hispanic children. However, African American, Native American / Pacific Islander, and Asian children were under represented. Multiracial and children of other races were over represented. This may be due to the particular states that were represented in the sample.

Analyses Related to the Construction of Scale Scores

Rasch scaling, the one parameter IRT model, was used to create ability estimates for each child on each construct and to examine the measurement properties of the information provided by each item. Data were analyzed using the Partial Credit Model (PCM; Masters, 1982), with Winsteps software (Linacre, 2012). A separate Rasch analysis was conducted for each of the six domains of development. 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 items do not share the same rating scales (i.e., use of the same number of rating scale categories and labels across items). In cases where each item has its own rating scale structure, the PCM is the appropriate model to apply. Specifically, 12 *GOLD*® items include a 0-9 scale, 6 items include a 0-11 scale, 16 items include a 0-13 scale, 25 items include a 0-15 scale, and one item includes a 0-19 scale. For each item, the 0 category represents "Not Yet" and the highest category represents abilities beyond the highest behavioral anchor.

Dimensionality

Rasch modeling assumes what is called unidimensionality, meaning that the items in question measure only one underlying latent construct. The unidimensionality of each scale was evaluated by using Mean Square (MNSQ) item 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 items (Bond & Fox, 2007). MNSQ values less than 2.0 can indicate that an item, though not fitting optimally with the measurement model, can still contribute useful information to the overall score on the measure. Items 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 item 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 item difficulties. They are also sensitive to unusual and unexpected item 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). These indexes were evaluated at all three time points.

Social Emotional Scale (9 items)

The PCAR showed that for the Social Emotional scale the Rasch dimension explained the majority of the variance in the data (fall = 72.5%, winter = 70.8%, spring = 74.7%) with the following eigenvalues: fall = 23.8, winter = 21.9, and spring = 26.5. The first contrast (the largest secondary dimension) had the following eigenvalues: fall = 1.6, winter = 1.6, and spring = 1.6. These secondary dimensions accounted for less than 5.2% of the unexplained variance (fall = 5.0%, winter = 5.1%, spring = 4.5%).

The fall fit statistics for all of the Social Emotional items were within acceptable limits. The fall infit MNSQ values ranged from 0.84 to 1.16. The outfit MNSQ values ranged from 0.82 to 1.26. The item total score correlations, with the item excluded from the total score, ranged from .66 to .77. The winter fit statistics for all of the Social Emotional items were within acceptable limits. The infit MNSQ values ranged from 0.85 to 1.49. Only one item had a MNSQ value between 1.4 and 2.0. The outfit MNSQ values ranged from 0.81 to 1.31. The item total score correlations, with the item excluded from the total score, ranged from .61 to .80. The spring fit statistics for all of the Social Emotional items were also within acceptable limits. The infit MNSQ values ranged from 0.86 to 1.63. Only one item had a MNSQ value above 1.4. The outfit MNSQ values ranged from 0.81 to 1.19. The item total score correlations, with the item excluded from the total score, ranged from .56 to .82.

Physical Scale (5 items)

The PCAR showed that for the Physical scale the Rasch dimension explained the majority of the variance in the data (fall = 73.6%, winter = 71.8%, spring = 74.3%) with the following eigenvalues: fall = 13.9, winter = 12.8, and spring = 14.4. The first contrast (the largest secondary dimension) had the following eigenvalues: fall = 2.2, winter = 2.2, and spring = 2.4. These secondary dimensions accounted for more than 11.0% of the unexplained variance (fall = 11.8%, winter = 12.3%, spring = 12.1%). Further investigation of the item loadings suggested that there is some evidence to support a gross motor scale (4, 5, and 6) and a fine motor scale (7.a and 7.b).

The fall fit statistics for all of the Physical items were within acceptable limits. The infit MNSQ values ranged from 0.86 to 1.25. The outfit MNSQ values ranged from 0.85 to 1.28. The item total score correlations, with the item excluded from the total score, ranged from .81 to .84. The winter fit statistics for all of the Physical items were within acceptable limits. The infit MNSQ values ranged from 0.86 to 1.30. The outfit MNSQ values ranged from 0.86 to 1.30. The item total score correlations, with the item excluded from the total score, ranged from .82 to .83. The spring fit statistics for all of the Physical items were also within acceptable limits. The infit MNSQ values ranged from 0.77 to 1.29. The outfit MNSQ values ranged from 0.77 to 1.35. The item total score correlations, with the item excluded from the total score, ranged from .82 to .85.

Language Scale (8 items)

The PCAR showed that for the Language scale the Rasch dimension explained the majority of the variance in the data (fall = 83.1%, winter = 81.2%, spring = 83.3%) with the following eigenvalues: fall = 39.3, winter = 34.6, and spring = 39.9. The first contrast (the largest secondary dimension) had the following eigenvalues: fall = 1.4, winter = 1.4, and spring = 1.4. These secondary dimensions accounted for less than 3.4% of the unexplained variance (fall = 2.9%, winter = 3.3%, spring = 3.0%).

The fall fit statistics for all of the Language items were within acceptable limits. The infit MNSQ values ranged from 0.82 to 1.50. Only one item had a MNSQ value above 1.4. The outfit MNSQ values ranged from 0.80 to 1.69. Only two items had a MNSQ value above 1.4. The item total score correlations, with the item excluded from the total score, ranged from .81 to .87. The winter fit statistics for all of the Language items were within acceptable limits. The infit MNSQ values ranged from 0.83 to 1.39. The outfit MNSQ values ranged from 0.83 to 1.33. The item total score correlations, with the item excluded from the total score, ranged from .83 to .87. The spring fit statistics for all of the Language items were also within acceptable limits. The infit MNSQ values ranged from 0.89 to 1.36. The outfit MNSQ values ranged from 0.87 to 1.34. The item total score correlations, with the item excluded from the total score, ranged from .84 to .89.

Cognitive Scale (10 items)

The PCAR showed that for the Cognitive scale the Rasch dimension explained the majority of the variance in the data (fall = 80.9%, winter = 80.9%, spring = 79.8%) with the following eigenvalues: fall = 42.2, winter = 42.4, and spring = 39.4. The first contrast (the largest secondary dimension) had the following eigenvalues: fall = 1.5, winter = 1.5, and spring = 1.5. These secondary dimensions accounted for less than 3.0% of the unexplained variance (fall = 2.9%, winter = 2.8%, spring = 2.9%).

The fall fit statistics for all of the Cognitive items were within acceptable limits. The infit MNSQ values ranged from 0.85 to 1.64. Only three items had MNSQ values between 1.4 and 2.0. The outfit MNSQ values ranged from 0.82 to 1.87. Only one item had a MNSQ value above 1.4. The item total score correlations, with the item excluded from the total score, ranged from .80 to .87. The winter fit statistics for all of the Cognitive items were within acceptable limits. The infit MNSQ values ranged from 0.85 to 1.21. The outfit MNSQ values ranged from 0.83 to 1.20. The

item total score correlations, with the item excluded from the total score, ranged from .84 to .90. The spring fit statistics for all of the Cognitive items were also within acceptable limits. The infit MNSQ values ranged from 0.83 to 1.20. The outfit MNSQ values ranged from 0.84 to 1.23. The item total score correlations, with the item excluded from the total score, ranged from .82 to .87.

Literacy Scale (16 items)

The PCAR showed that for the Literacy scale the Rasch dimension explained the majority of the variance in the data (fall = 79.6%, winter = 83.8%, spring = 86.3%) with the following eigenvalues: fall = 62.5, winter = 82.8, and spring = 109.9. The first contrast (the largest secondary dimension) had the following eigenvalues: fall = 1.5, winter = 1.5, and spring = 1.5. These secondary dimensions accounted for less than 3.1% of the unexplained variance (fall = 3.0%, winter = 2.7%, spring = 2.2%).

The fall fit statistics for all but one of the Literacy items were within acceptable limits. The infit MNSQ values ranged from 0.75 to 3.00. Only one item had a MNSQ value between 1.4 and 2.0 and one item (19.b) had a MNSQ item above 2.0. The outfit MNSQ values ranged from 0.73 to 4.07. Only one item (19.b) had a MNSQ value above 1.4. The item total score correlations, with the item excluded from the total score, ranged from .47 to .80. The winter fit statistics for all but one of the Literacy items were within acceptable limits. The infit MNSQ values ranged from 0.81 to 2.71. Only one item had a MNSQ value between 1.4 and 2.0 and one item (19.b) had a value above 2.0. The outfit MNSQ values ranged from 0.81 to 5.23. Only one item had a MNSQ value between 1.4 and 2.0 and one item (19.b) had a value above 2.0. The item total score correlations, with the item excluded from the total score, ranged from .59 to .80. The spring fit statistics for all but one of the Literacy items were also within acceptable limits. The infit MNSQ values ranged from 0.79 to 2.91.

Only one item (19.b) had a MNSQ value above 1.4. The outfit MNSQ values ranged from 0.79 to 9.90. Only one item (19.b) had a MNSQ value above 1.4. The item total score correlations, with the item excluded from the total score, ranged from .58 to .81.

Mathematics Scale (12 items)

The PCAR showed that for the Mathematics scale the Rasch dimension explained the majority of the variance in the data (fall = 74.0%, winter = 75.8%, spring = 79.7%) with the following eigenvalues: fall = 34.2, winter = 37.5, and spring = 47.0. The first contrast (the largest secondary dimension) had the following eigenvalues: fall = 2.2, winter = 2.3, and spring = 2.3. These secondary dimensions accounted for less than 4.8% of the unexplained variance (fall = 4.7%, winter = 4.6%, spring = 3.9%).

The fall fit statistics for all of the Mathematics items were within acceptable limits. The infit MNSQ values ranged from 0.77 to 1.67. Only two items had MNSQ values between 1.4 and 2.0. The outfit MNSQ values ranged from 0.74 to 1.78. Only two items had MNSQ values between 1.4 and 2.0. The item total score correlations, with the item excluded from the total score, ranged from .51 to .74. The winter fit statistics for all of the Mathematics items were within acceptable limits. The infit MNSQ values ranged from 0.81 to 1.43. Only one item had a MNSQ value above 1.4. The outfit MNSQ values ranged from 0.85 to 1.54. Only one item had a MNSQ value above 1.4. The item total score correlations, with the item excluded from the total score, ranged from .60 to .71. The spring fit statistics for all but one of the Mathematics items were also within acceptable limits. The infit MNSQ values ranged from 0.77 to 2.80. Only one item (22.b) had a MNSQ value above 1.4. The outfit MNSQ values ranged from 0.72 to 2.80. Only one item (22.b) had a MNSQ value above 1.4. The item total score correlations, with the item excluded from the total score, ranged from .64 to .74.

In summary, with the few exceptions noted above, 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.

Rating Scale Category Effectiveness

Given that this report focuses on only one age or color band, it is important to examine whether the teachers used each entire rating scale when making placements on the developmental progressions. This is of course not an issue when using a large sample that includes children across all age or color bands. If children from birth to third grade were included in the sample, it would be very reasonable to expect that the entire rating scale would be used for each item. Kindergarten children are close to the middle of the intended age range (birth to third grade) for *GOLD*® and so it is possible that teachers used the entire rating scales for each progression. The ranges of rating scale points used was compared to the full range of scale points for each item and at each time point. This was done to evaluate if it is reasonable to apply Rasch modeling to the data. It is recommended that for each item, each rating scale category is assigned to a minimum of 10 children. The use of rating scale categories was also examined to provide information about whether teachers utilized the instrument in the manner in which it was intended.

Rating scale category effectiveness is a measure of validity for the items. The median score for each item was examined to determine if it increased over time as expected. The average of the ability estimates, based on the total item scores, for all persons 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). Thresholds (also called step calibrations) are the difficulty levels estimated for choosing 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. 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). These plots indicate the probability of a child being placed on a particular response category or level of each developmental progression given their overall ability or total measure score for the associated scale.

As can be seen from Table 2, teachers used more of the rating scale categories in the spring of the year than they did at the other two assessment time points. They also tended to use the entire rating scale or almost the entire rating scales in the spring. This indicates that as kindergarten children grow and develop they are much more likely to be placed at the upper ends of the developmental progressions in the spring than in the fall. Therefore, the results of these analyses will be reported in detail for the spring assessment time point.

Social Emotional Scale

Eight of the nine items include a scale that ranges from 0 to 13. One of the items includes a scale that ranges from 0 to 11. In the spring, the teachers used the entire rating scale to place kindergarten children on the progressions for 8 of the 9 items. For one of the items (1.c) the teachers in this sample used all but the highest category. The median placements for all items advanced over time as expected. The median placements ranged from 5 to 7 in fall, 6 to 8 in the winter, and 7 to 8 in the spring. The observed sample averages for each response category generally advanced as expected across the rating scales. For 4 of the 9 items there were no disordered averages. There was one disordered average for 4 of the 9 items and 2 disordered averages for 1 of the 9 items. The thresholds generally advanced for each response category as expected across the

rating scales. However, for 1 of the 9 items there were two disordered thresholds. There were three disordered thresholds for 7 of the 9 items and four disordered thresholds for 1 of the 9 items.

Physical Scale

Three of the five items include a scale that ranges from 0 to 13. Two of the items include a scale that ranges from 0 to 15. In the spring, the teachers used the entire rating scale to place kindergarten children on the progressions for all 5 items. The median placements for all items advanced over time as expected. The median placements were 7 in fall, 8 in the winter, and 8 to 9 in the spring. The observed sample averages for each response category generally advanced as expected across the rating scales. For 2 of the 5 items there was one disordered average. There were two disordered averages for 3 of the 5 items. The thresholds generally advanced for each response category as expected across the rating scales. However, for 1 of the 5 items there were two disordered thresholds. There were three disordered thresholds for 4 of the 5 items.

Language Scale

Six of the eight items include a scale that ranges from 0 to 15. One of the items includes a scale that ranges from 0 to 11. One of the items includes a scale that ranges from 0 to 13. In the spring, the teachers used the entire rating scale to place kindergarten children on the progressions for all of 8 items. The median placements for all items advanced over time as expected. The median placements ranged from 6 in the fall, 6 to 7 in the winter, and 7 to 8 in the spring. The observed sample averages for each response category generally advanced as expected across the rating scales. For 3 of the 8 items there were no disordered averages. There was one disordered average for 4 of the 8 items and 2 disordered averages for 1 of the 8 items. The thresholds generally advanced for each response category as expected across the rating scales. There were no disordered thresholds for

2 of the 8 items. However, for 3 of the 8 items there were two disordered thresholds. There were three disordered thresholds for 2 of the 8 items and four disordered thresholds for 1 of the 8 items.

Cognitive Scale

Six of the ten items include a scale that ranges from 0 to 15. Four of the items include a scale that ranges from 0 to 13. In the spring, the teachers used the entire rating scale to place kindergarten children on the progressions for 8 of the 10 items. For two of the items (12.a and 14.b) the teachers in this sample used all but the highest category. The median placements for all items advanced over time as expected. The median placements ranged from 5 to 6 in fall, 6 to 7 in the winter, and 7 to 8 in the spring. The observed sample averages for each response category generally advanced as expected across the rating scales. For 7 of the 10 items there were no disordered averages. There were 3 items with one disordered average. The thresholds generally advanced for each response category as expected across the rating scales. There were two items with no disordered thresholds. However, for 2 of the 8 items there was one disordered threshold. There were two disordered thresholds for 2 of the 8 items and three disordered thresholds for 4 of the 8 items.

Literacy Scale

Seven of the sixteen items include a scale that ranges from 0 to 9. Three of the items include a scale that ranges from 0 to 11. Five of the items include a scale that ranges from 0 to 15. One item includes a scale that ranges from 0 to 19. In the spring, the teachers used the entire rating scale to place kindergarten children on the progressions for 15 of the 16 items. For one of the items (19.b) the teachers in this sample used all but the highest 3 categories. The median placements for all items advanced over time as expected. The median placements ranged from 0 to 12 in fall, 1 to 13 in the winter, and 2 to 14 in the spring. The observed sample averages for each response category generally

advanced as expected across the rating scales. For 8 of the 16 items there were no disordered averages. There was one disordered average for 5 of the 16 items, 2 disordered averages for 1 of the 16 items, and 3 disordered categories for 2 of the 16 items. The thresholds generally advanced for each response category as expected across the rating scales. However, for 4 of the 16 items there were two disordered thresholds. There were three disordered thresholds for 5 of the 16 items and four disordered thresholds for 6 of the 16 items. For one of the items (19.b) there were 6 disordered thresholds.

Mathematics Scale

Four of the twelve items include a scale that ranges from 0 to 9. One of the items includes a scale that ranges from 0 to 11. One of the items includes a scale that ranges from 0 to 13. Six of the items includes a scale that ranges from 0 to 15. In the spring, the teachers used the entire rating scale to place kindergarten children on the progressions for 5 of the 12 items. For one of the items the teachers in this sample used all but the highest category. For one of the items the teachers in this sample used all but the highest two categories. For four of the items the teachers in this sample used all but the highest three categories. For one of the items the teachers in this sample used all but the highest four categories. The median placements for all items advanced over time as expected. The median placements ranged from 0 to 6 in fall, 0 to 7 in the winter, and 2 to 8 in the spring. The observed sample averages for each response category generally advanced as expected across the rating scales. For 3 of the 12 items there were no disordered averages. There was one disordered average for 5 of the 12 items, two disordered averages for 3 of the 12 items, and four disordered averages for 1 of the 12 items. The thresholds generally advanced for each response category as expected across the rating scales. There were no disordered thresholds for 2 of the 12 items. However, for 1 of the 12 items there was one disordered thresholds. There were two disordered

thresholds for 1 of the 12 items, three disordered thresholds for 4 of the 12 items, and four disordered thresholds for 3 of the 12 items. For one item (22.a) there were seven disordered thresholds across a 0 to 15 scale.

These results do suggest potential issues related to using some of the developmental progressions, as noted, with kindergarten children. However, these results should be taken with caution. They are certainly due in substantial part to applying the Rasch measurement model using a sample consisting of only one age or color band. These issues have not been observed with previous samples that included large, nationally representative sample from all of the age or color bands across the intended age for the measure.

Item Difficulty Measures

For all six scales, the item location hierarchy appeared to be generally consistent with the expected developmental trajectory for typically developing kindergarten children. Tables 4 through 9 list the item difficulty estimates from highest to lowest along with the standard errors for these estimates and the associated fit statistics. These results are evaluated for the fall assessment time period given that *GOLD*® is most commonly used with Kindergarten children for kindergarten entry assessment.

For the Social Emotional Scale, the item pertaining to a child's ability to form relationships with adults (2.a) was estimated as the easiest item (-1.82). The item pertaining to a child's ability to solve social problems (3.b) was found to be the most difficult item (1.26). The range of both item difficulties (-1.82 to 1.26) and item rating scale anchor point locations (-7.35 to 9.82) was considered wide enough for reasonable separation of children according to underlying ability.

For the Physical Scale, the item pertaining to a child's ability to demonstrate traveling skills (4) was estimated as the easiest item (-1.00). The item pertaining to a child's ability to demonstrate gross

motor manipulative skills (6) was found to be the most difficult item (.76). The range of overall item difficulties (-1.00 to .76) and item rating scale anchor point locations (-9.29 to 12.23), although narrower that for the other scales and based on fewer items, was considered wide enough for reasonable separation of children according to underlying ability.

For the Language Scale, the item pertaining to a child's ability to comprehend language (8.a) was found to be the easiest item (-1.50). The item pertaining to a child's ability to tell about another time or place (9.d) was estimated as the most difficult item. The range of item difficulties (-1.50 to 1.59) and item rating scale anchor point locations (-9.92 to 13.42) was considered wide enough for reasonable separation of children according to underlying ability.

For the Cognitive Scale, the item pertaining to a child's ability to think symbolically (14.a) was found to be the easiest item (-2.07). The items pertaining to a child's ability to show flexibility and inventiveness in thinking (11.e) was estimated as the most difficult item (2.28). The range of overall item difficulties (-2.07 to 2.28) and item rating scale anchor point locations (-8.52 to 12.19) was considered sufficient for separation of children across the range of underlying abilities.

For the Literacy Scale, the item pertaining to a child's ability to write their name (19.a) was estimated as the easiest item (-2.71). The item pertaining to a child's ability to read fluently (18.e) was found to be the most difficult item (1.65). The range of both item difficulties (-2.71 to 1.65) and item rating scale anchor point locations (-5.43 to 6.32) was considered wide enough for reasonable separation of children according to underlying ability.

For the Mathematics Scale, the item pertaining to a child's ability to explore shapes (21.b) was estimated to be the easiest item (-1.96). The item pertaining to a child's ability to understand and use place value and base ten (20.d) was found to be the most difficult item (2.20). The range of both item difficulties (-1.96 to 2.20) and item rating scale anchor point locations (-7.36 to 8.82) was considered wide enough for reasonable separation of children according to underlying ability.

In summary, the developmental pathway that is formed for each scale indicates a progression from the easiest to the most difficult items that generally aligns with expectations from developmental theory. In addition, the range of difficulties for each scale is the widest that has been observed from use with kindergarten children, suggesting that kindergarten teachers in the field are able to separate children according to their analysis of appropriate evidences collected. It is also important to recognize, as indicated, that the range of item 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 and person reliabilities were evaluated using both sample-based and model-based coefficients. Each of these indexes was evaluated for the fall, winter, and spring assessment periods. 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 separation index indicates an estimate in standard error units of the spread or separation of items 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 reliability means that there is a high probability of replicating the same separation of persons or items across measurements. Specifically, person separation reliability estimates the replicability of person placement across other items measuring the same construct. Similarly, item separation reliability estimates the replicability of item placement along the construct development pathway if the same items were given to another sample with similar ability levels. The person reliability

provided is similar to the classical or traditional test reliability whereas the item reliability has no classical equivalent. Low values in person and item reliability may indicate a narrow range of person or item measures. It may also indicate that the number of items 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.

Table 3 contains the reliability coefficients from the information yielded by each of the scale scores. Across all domains of development and for all three time points, all of the item reliability values, both sample-based and model-based, were greater than .99. Therefore these values are not reported in the table. Similarly, all of the item separation indexes for all domains of development were very high and are therefore not included in the table. Specifically, for the Social Emotional scale scores, all of the item separation indexes across all three time points, both sample-based and model-based, were greater than 47. For the Physical scale scores, all of the item separation indexes across all three time points, both sample-based and model-based, were greater than 50. For the Language scale scores, all of the item separation indexes across all three time points, both sample-based and model-based, were greater than 42. For the Cognitive scale scores, all of the item separation indexes across all three time points, both sample-based and model-based, were greater than 29. For the Literacy scale scores, all of the item separation indexes across all three time points, both sample-based and model-based, were greater than 90. For the Mathematics scale scores, all of the item separation indexes across all three time points, both sample-based and model-based, were greater than 66. The person based reliability coefficients are outlined below by domain of development. Taken together, these findings indicate it is reasonable to expect very highly consistent estimates of item difficulty levels across samples.

Social Emotional Scale

Based on the Rasch reliability indexes, the scale scores appear to yield adequately reliable information from this sample, as evidenced by sample-based person separation indexes that ranged from 2.63 to 2.98, and model-based person separation indexes that ranged from 3.02 to 3.49. Similarly, the sample-based person reliability indexes ranged from .87 to .90 and the model-based person reliabilities ranged from .90 to .92. Cronbach's alpha values ranged from .89 to .91 indicating adequate internal consistency reliability.

Physical Scale

Based on the Rasch reliability indexes, the scale scores appear to adequately reliable information from this sample, as evidenced by sample-based person separation indexes that ranged from 2.51 to 2.66, and model-based person separation indexes that ranged from 3.08 to 3.17. Similarly, the sample-based person reliability indexes ranged from .86 to .88 and the model-based person reliabilities ranged from .90 to .91. Cronbach's alpha values ranged from .89 to .90 indicating adequate internal consistency reliability.

Language Scale

Based on the Rasch reliability indexes, the scale scores appear to yield highly reliable information from this sample, as evidenced by sample-based person separation indexes that ranged from 3.10 to 3.44, and model-based person separation indexes that ranged from 3.66 to 4.00. Similarly, the sample-based person reliability indexes ranged from .90 to .92 and the model-based person reliabilities ranged from .93 to .94. Cronbach's alpha values ranged from .95 to .96 indicating high internal consistency reliability.

Cognitive Scale

Based on the Rasch reliability indexes (see Table 3), the scale scores appear to yield highly reliable information from this sample, as evidenced by sample-based person separation indexes that ranged from 3.00 to 3.80, and model-based person separation indexes that ranged from 3.48 to 4.44. Similarly, the sample-based person reliability indexes ranged from .90 to .94 and the model-based person reliabilities ranged from .92 to .95. Cronbach's alpha values ranged from .96 to .97 indicating high internal consistency reliability.

Literacy Scale

Based on the Rasch reliability indexes, the scale scores appear to yield highly reliable information from this sample, as evidenced by sample-based person separation indexes that ranged from 3.10 to 3.34, and model-based person separation indexes that ranged from 3.61 to 3.92. Similarly, the sample-based person reliability indexes ranged from .91 to .92 and the model-based person reliabilities ranged from .93 to .94. Cronbach's alpha values ranged from .90 to .92 indicating high internal consistency reliability.

Mathematics Scale

Based on the Rasch reliability indexes, the scale scores appear to adequately reliable information from this sample, as evidenced by sample-based person separation indexes that ranged from 2.17 to 2.28, and model-based person separation indexes that ranged from 2.49 to 2.70. Similarly, the sample-based person reliability indexes ranged from .82 to .84 and the model-based person reliabilities ranged from .86 to .88. Cronbach's alpha values ranged from .87 to .89 indicating adequate internal consistency reliability.

In summary, these results indicate that it is reasonable to expect highly reliable estimates of child ability levels when using *GOLD*® with kindergarten children for three of the domains of

development: Cognitive, Language, and Literacy development. All of the Rasch reliability indexes were greater than .90 for these scales across all three time points and all the person separation indexes were greater than 3.00. These results also indicate that it is reasonable to expect adequately reliable estimates of child ability levels when using *GOLD*® with kindergarten children for the remaining three domains of development: Mathematics, Physical, and Social Emotional development. All of the Rasch reliability indexes were greater than .80 for these scales across all three time points and all the person separation indexes were greater than 2.00. It is also important to note that for all of the indexes of reliability the values were the highest for the spring assessment period suggesting that as kindergarten teachers know the children more completely and have a wider range of evidences from the classroom to analyze when making placements on the developmental progressions, they also exhibit higher reliability indexes.

Scale Scores and Widely Held Expectations

Previous research, using a nationally representative sample from all 50 states and including all age / color bands for which *GOLD*® was designed and validated was used to calibrate scale scores. The Rasch PCM was used to create a raw score to scale conversion process. The scale scores have been scaled to conform to a distribution with a mean of 500 and a range from 0 to 1,000. The scale scores for the children in the current sample were created by first calculating raw scores or simple sums of the placements on each progression within each domain of development. If a child did not have complete rating scale data, but was rated by the teacher on at least 70% of the items on a respective scale, then the child's scale mean rating, rescaled to accommodate the different scaling across the missing items, was substituted for each of the missing ratings. The scale scores were created by transforming the raw scores into interval level Rasch ability estimates for each child.

For each scale score and age / grade band, as shown in Tables 10, the scale mean, standard deviation, and quartile boundaries are reported for each of the three checkpoints. The same information is also provided for fall to spring gains. The standard errors of measurement (SEM) are reported at the scale mean for each respective time point. In all IRT models, unlike with classical measurement models, the SEM can be estimated for each scale score point. These results highlight the fact that scale scores from *GOLD*® are sensitive to child growth over time. They are useful to teachers who are interested in tracking children's growth over time on an interval scale, gathering a general sense of the overall domains of development that are within expected ranges, and are attempting to gain a broad picture of a child's overall strengths and areas for growth and support.

While scale scores can be useful to teachers, and can also give some interpretation guidelines for comparison when working with aggregated mean scores from classrooms of children, Widely Held Expectations (WHE) scores illustrate how *GOLD*® provides criterion referenced interpretation guidelines for teachers. These scores are available at the item and scale score levels and indicate when a child is below, meeting, or exceeding developmental criteria for their age group that are supported by research, expert opinion, and developmental theory. At the item level, these scores provide teachers with very useful guidance about child progress and instructional support. Although it is desirable, it is often very difficult for teachers to differentiate instruction for every child across every instructional objective and domain of development. WHE scores are very useful to guide teachers toward each child's areas of need. Table 11 shows the percentage of children below, meeting, and exceeding WHE for each time point.

Figure 1 displays these same results graphically as trends for WHE scores over time. For all six domains of development, the percentage of children in the "Below" category declined over the course of the academic year. For all six domains of development, the percentage of children in the "Exceeds" category increased over time. For all but two of the domains of development, the

"Meets" category increased over time as well. For Literacy, the "Meets" category declined from winter to spring due to the percentage of children in the "Exceeds" category expanding greatly. Similarly, for Mathematics, the "Meets" category declined from fall to spring due to the fact that the percentage of children in the "Exceeds" category expanded greatly.

Table 12 shows the correlations between the scale scores across the assessment time periods. For ease of reading the values across the rows, redundant information is included in the table. For each of the scales, the smallest correlation values were found for the association between fall and spring scores. The strongest correlation values were found for the association between winter and spring scores, and these values were very similar to the values for the association between fall and winter scores. These results indicate, as expected, that assessment scores closest in time are most highly correlated and that as a teacher gets to know a child more over the course of the academic year and has a richer portfolio of evidences to support placements on the developmental progressions, the correlations are higher.

Summary

Overall, *GOLD®* appears to continue to yield highly reliable scores as indicated by both the classical and Rasch reliability statistics. The high reliability statistics were not only found in this sample, but are similar to those found in earlier nationally representative normative studies. This particularly noteworthy considering this study only included one age group, kindergarten children. The results also demonstrate strong statistical evidence that the items within each scale generally work very well together to measure a single underlying construct or domain of development. The items within each scale yield information that fits the statistical model that was used to develop the scoring strategy that is used to create the scale scores. The results further demonstrate evidence that the ratings can be successfully organized by developmental domain or latent construct generally as

intended by the instrument development team. Analyses of the dimensionality of each scale score strongly suggest that *GOLD*® ratings measure six distinct domains of development and that each satisfies the Rasch model assumption of unidimensionality. The model fit statistics suggest that the data are a good fit for the Rasch rating scale model.

There is also statistical evidence that teachers are able to use the rating scale to place children along a progression of development and learning. When the items within each domain of development are arranged from the easier objectives for children to master to the most difficult objectives for children to master, the hierarchy that is created matches very well with what developmental theory indicates. Therefore, the range of item difficulties indicates that each section of *GOLD*® can be used by teachers to help them understand the developmental trajectory that most children will follow.

Data from a wider range of developmental levels is needed to make firm conclusions, especially given the relatively smaller numbers of children placed at the upper and lower ends of the rating scale progressions. Using a sample with nationally representative samples of children at all age levels, particularly those at the ends of the intended age range, is needed to reevaluate rating scale effectiveness.

Future research could focus on further measures of the degree of association between *GOLD*® scale scores and external measures of child developmental progress. It would also be helpful to conduct additional inter rater reliability studies. These studies can focus on both procedural fidelity and agreement with expert raters as well as variance decomposition methods that address generalizability. As teachers around the country gain more experience and training with the use of the measure, it may also be helpful to conduct studies that examine the proportion of the variability in ratings that is between and within raters, the sensitivity of the scores to growth over time, and continuing examination of the differences between subgroups of children. In addition,

future research is needed to evaluate whether teachers are collecting sufficient quantity of high quality, valid evidences of child growth, development, and learning to support placements on the developmental progressions *in GOLD*®.

References

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.

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.

Kim, D-H., Lambert, R. G., & Burts, D. C. (2013). Evidence of the validity of *Teaching Strategies GOLD®* assessment tool for English language learners and children with disabilities. *Early Education and Development, 24*, 574-595.

Kim, D-H., Lambert, R. G., & Burts, D. C. (2014). Validating a developmental scale for young children using the Rasch Model: Applicability of the *Teaching Strategies GOLD®*assessment system. *Journal of Applied Measurement, 15(4), 405-421.*

Lambert, R. & Kim, D-H., Taylor, H., & McGee, J. (2010). *Technical manual for the Teaching Strategies GOLD assessment system.* Technical Report. Charlotte, N.C.: *Center for Educational Measurement and Evaluation*, University of North Carolina Charlotte.

Lambert, R. G., Kim, D-H., & Burts, D. C. (2012). [The measurement properties of the *Teaching Strategies GOLD®* assessment system]. Unpublished raw data.

Lambert, R. G., Kim, D-H., & Burts, D. C. (2013). Using teacher ratings to track the growth

and development of young children using the *Teaching Strategies GOLD®* assessment system. *Journal of Psychoeducational Assessment*. doi:0734282913485214

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

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

Table 1
Demographic characteristics of the sample

Characteristic	Percent
White (non-Hispanic)	49.6
African American (non-Hispanic)	8.2
Native American / Pacific Islander (non-Hispanic)	0.7
Asian (non-Hispanic)	2.6
Multirace / other (non-Hispanic)	16.5
Hispanic	22.4
Male	49.1
Female	50.9
English	76.6
Spanish	19.9
Other	3.5
Has an IEP	1.3
Qualifies for NSLP	6.7

Table 2
Use of the full range of rating scale points by item and assessment period

Scale	Item	Scaling of Full Progression Min	Scaling of Full Progression Max	Fall Placements Min	Fall Placements Max	Fall Placements Median	Winter Placements Min	Winter Placements Max	Winter Placements Median	Spring Placements Min	Spring Placements Max	Spring Placements Median
Social Emotional	1.A	0	13	0	12	6	0	12	7	0	13	8
	1.B	0	13	0	13	6	0	12	7	0	13	8
	1.C	0	13	0	12	7	0	12	8	0	12	8
	2.A	0	11	0	10	7	0	11	8	0	11	8
	2.B	0	13	0	12	6	0	12	7	0	13	8
	2.C	0	13	0	12	6	0	13	7	0	13	8
	2.D	0	13	0	10	6	0	12	7	0	13	8
	3.A	0	13	0	12	6	0	12	7	0	13	8
	3.B	0	13	0	13	5	0	12	6	0	13	7
Physical	4	0	13	0	12	7	0	13	8	0	13	8
	5	0	15	0	14	7	0	14	8	0	15	9
	6	0	15	0	14	7	0	14	8	0	15	8
	7.A	0	13	0	13	7	0	13	8	0	13	8
	7.B	0	13	0	12	7	0	12	8	0	13	8
Language	8.A	0	15	0	10	6	0	11	7	0	15	8
	8.B	0	11	0	10	6	0	11	7	0	11	8
	9.A	0	15	0	14	6	0	14	7	0	15	7
	9.B	0	15	0	11	6	0	11	7	0	15	8
	9.C	0	15	0	14	6	0	14	7	0	15	8
	9.D	0	13	0	12	6	0	12	6	0	13	7
	10.A	0	15	0	14	6	0	15	7	0	15	8
	10.B	0	15	0	12	6	0	14	7	0	15	8
Cognitive	11.A	0	15	0	12	6	0	12	6	0	15	8
	11.B	0	13	0	12	6	0	12	6	0	13	7
	11.C	0	13	0	12	6	0	13	6	0	13	7
	11.D	0	15	0	14	6	0	15	6	0	15	7
	11.E	0	15	0	14	5	0	13	6	0	15	7
	12.A	0	15	0	8	5	0	11	6	0	14	7
	12.B	0	15	0	14	6	0	13	6	0	15	7
	13	0	13	0	8	5	0	10	6	0	13	7
	14.A	0	13	0	8	6	0	10	6	0	13	8
	14.B	0	15	0	9	6	0	10	7	0	14	8
Literacy	15.A	0	11	0	10	5	0	11	6	0	11	8
	15.B	0	9	0	9	4	0	9	7	0	9	8
	15.C	0	15	0	15	4	0	15	6	0	15	8
	15.D	0	11	0	11	1	0	10	3	0	11	4
	16.A	0	9	0	9	6	0	9	7	0	9	8
	16.B	0	9	0	9	3	0	9	5	0	9	7
	17.A	0	15	0	12	5	0	13	6	0	15	7
	17.B	0	11	0	11	4	0	11	6	0	11	8
	18.A	0	15	0	14	4	0	14	5	0	15	7
	18.B	0	9	0	9	5	0	9	7	0	9	8
	18.C	0	15	0	14	4	0	15	5	0	15	7
	18.D	0	9	0	5	0	0	9	1	0	9	2
	18.E	0	9	0	9	0	0	8	1	0	9	2
	19.A	0	15	0	15	12	0	15	13	0	15	14
	19.B	0	19	0	19	8	0	14	10	0	16	11
	19.C	0	9	0	6	0	0	8	1	0	9	2
Mathematics	20.A	0	15	0	12	5	0	13	7	0	14	8
	20.B	0	15	0	14	5	0	13	5	0	15	8
	20.C	0	15	0	14	6	0	12	7	0	15	8
	20.D	0	9	0	7	0	0	5	0	0	6	2
	20.E	0	9	0	3	0	0	6	1	0	6	2
	20.F	0	9	0	6	0	0	7	1	0	9	2
	21.A	0	9	0	9	6	0	9	6	0	9	7
	21.B	0	15	0	9	5	0	12	6	0	12	7
	22.A	0	15	0	11	5	0	14	6	0	15	7
	22.B	0	13	0	7	4	0	8	4	0	10	5
	22.C	0	11	0	8	2	0	8	3	0	9	4
	23	0	15	0	10	6	0	10	6	0	11	8

Table 3
Reliability coefficients for all scales across the three assessment periods

Domain	Index	Fall	Winter	Spring
Social Emotional	Cronbach's alpha	.89	.90	.91
	Sample-based Person Separation Index	2.63	2.85	2.98
	Sample-based Person Reliability	.87	.89	.90
	Model-based Person Separation Index	3.02	3.31	3.49
	Model-based Person Reliability	.90	.92	.92
Physical	Cronbach's alpha	.90	.89	.90
	Sample-based Person Separation Index	2.51	2.46	2.66
	Sample-based Person Reliability	.86	.86	.88
	Model-based Person Separation Index	3.08	2.96	3.17
	Model-based Person Reliability	.90	.90	.91
Language	Cronbach's alpha	.96	.95	.96
	Sample-based Person Separation Index	3.10	3.06	3.44
	Sample-based Person Reliability	.91	.90	.92
	Model-based Person Separation Index	3.60	3.56	4.00
	Model-based Person Reliability	.93	.93	.94
Cognitive	Cronbach's alpha	.96	.96	.97
	Sample-based Person Separation Index	3.00	3.41	3.80
	Sample-based Person Reliability	.90	.92	.94
	Model-based Person Separation Index	3.48	3.97	4.44
	Model-based Person Reliability	.92	.94	.95
Literacy	Cronbach's alpha	.90	.91	.92
	Sample-based Person Separation Index	3.10	3.18	3.34
	Sample-based Person Reliability	.91	.91	.92
	Model-based Person Separation Index	3.61	3.72	3.92
	Model-based Person Reliability	.93	.93	.94
Mathematics	Cronbach's alpha	.87	.87	.89
	Sample-based Person Separation Index	2.22	2.17	2.28
	Sample-based Person Reliability	.83	.82	.84
	Model-based Person Separation Index	2.53	2.49	2.70
	Model-based Person Reliability	.86	.86	.88

Table 4
Fall item level statistics and difficulty estimates for the Cognitive scale

	Item	Item Difficulty	SE	Infit Mnsq	Outfit Mnsq	Item-Measure *r* Item Included	Item-Measure *r* Item Excluded
Cognitive	11.E	2.28	0.01	1.00	1.01	0.84	0.84
	11.B	1.74	0.01	0.96	0.98	0.84	0.83
	12.B	1.24	0.01	0.91	0.85	0.83	0.81
	11.C	1.09	0.01	0.88	0.89	0.85	0.83
	11.A	0.67	0.03	1.54	1.31	0.80	0.85
	11.D	-0.26	0.01	0.85	0.82	0.82	0.80
	14.B	-1.28	0.03	1.55	1.36	0.78	0.82
	12.A	-1.59	0.03	1.03	1.02	0.85	0.85
	13	-1.81	0.03	1.64	1.87	0.81	0.87
	14.A	-2.07	0.03	1.19	1.32	0.83	0.86

Table 5
Fall item level statistics and difficulty estimates for the Language scale

	Item	Item Difficulty	SE	Infit Mnsq	Outfit Mnsq	Item-Measure *r* Item Included	Item-Measure *r* Item Excluded
Language	9.D	1.59	0.01	0.91	0.91	0.87	0.87
	9.A	1.29	0.01	1.03	1.01	0.85	0.85
	10.A	1.03	0.01	0.91	0.89	0.86	0.85
	10.B	0.59	0.03	1.50	1.44	0.84	0.87
	9.B	-0.74	0.03	1.40	1.69	0.75	0.81
	9.C	-0.97	0.01	0.82	0.80	0.87	0.85
	8.B	-1.29	0.03	1.27	1.23	0.83	0.86
	8.A	-1.50	0.04	1.25	1.16	0.84	0.86

Table 6

Fall item level statistics and difficulty estimates for the Literacy scale

	Item	Item Difficulty	SE	Infit Mnsq	Outfit Mnsq	Item-Measure *r* Item Included	Item-Measure *r* Item Excluded
Literacy	18.E	1.65	0.03	0.84	0.74	0.55	0.47
	19.C	1.52	0.03	0.98	0.99	0.51	0.47
	18.D	1.51	0.03	0.93	1.19	0.52	0.47
	18.C	1.27	0.01	0.86	0.88	0.77	0.73
	15.C	1.08	0.01	0.87	0.88	0.80	0.76
	15.D	1.03	0.02	1.01	1.03	0.64	0.65
	18.A	0.93	0.01	0.75	0.73	0.81	0.73
	19.B	0.30	0.01	3.00	4.07	0.60	0.77
	16.B	-0.23	0.01	0.98	0.98	0.76	0.75
	17.A	-0.30	0.01	0.84	0.86	0.72	0.67
	17.B	-0.47	0.01	0.76	0.74	0.84	0.77
	15.A	-1.20	0.01	1.13	1.21	0.73	0.77
	15.B	-1.27	0.01	1.02	1.04	0.80	0.80
	18.B	-1.53	0.01	0.87	0.87	0.82	0.79
	16.A	-1.56	0.01	1.35	1.38	0.72	0.78
	19.A	-2.71	0.01	1.46	1.39	0.60	0.70

Table 7
Fall item level statistics and difficulty estimates for the Mathematics scale

	Item	Item Difficulty	SE	Infit Mnsq	Outfit Mnsq	Item-Measure *r* Item Included	Item-Measure *r* Item Excluded
Mathematics	20.D	2.20	0.04	0.97	1.20	0.49	0.51
	20.F	1.85	0.03	1.04	1.30	0.52	0.54
	20.E	0.97	0.03	0.88	0.96	0.59	0.55
	20.B	0.67	0.01	0.77	0.74	0.78	0.73
	20.C	0.53	0.01	0.94	0.91	0.73	0.73
	22.C	0.42	0.01	1.43	1.45	0.59	0.70
	22.A	0.34	0.01	1.15	1.18	0.70	0.72
	22.B	-1.06	0.01	1.67	1.78	0.62	0.74
	20.A	-1.17	0.01	1.02	1.02	0.72	0.74
	21.A	-1.17	0.01	0.83	0.81	0.69	0.65
	23	-1.61	0.01	1.16	1.26	0.64	0.67
	21.B	-1.96	0.01	0.90	0.91	0.72	0.69

Table 8
Fall item level statistics and difficulty estimates for the Physical scale

	Item	Item Difficulty	SE	Infit Mnsq	Outfit Mnsq	Item-Measure *r* Item Included	Item-Measure *r* Item Excluded
Physical	6	0.76	0.01	0.89	0.89	0.84	0.82
	5	0.72	0.01	1.00	0.97	0.83	0.81
	7.A	0.11	0.01	1.00	1.03	0.82	0.82
	7.B	-0.60	0.01	1.25	1.28	0.79	0.84
	4	-1.00	0.01	0.86	0.85	0.84	0.81

Table 9

Fall item level statistics and difficulty estimates for the Social Emotional scale

	Item	Item Difficulty	SE	Infit Mnsq	Outfit Mnsq	Item-Measure *r* Item Included	Item-Measure *r* Item Excluded
Social Emotional	3.B	1.26	0.01	1.00	1.00	0.77	0.76
	3.A	0.92	0.01	0.90	0.85	0.77	0.74
	2.B	0.81	0.01	0.88	0.81	0.76	0.73
	2.C	0.41	0.01	0.84	0.82	0.79	0.77
	1.A	-0.09	0.01	1.16	1.18	0.72	0.76
	1.B	-0.14	0.02	1.16	1.26	0.70	0.75
	1.C	-0.17	0.01	1.07	1.07	0.73	0.75
	2.D	-1.18	0.02	1.21	1.20	0.67	0.74
	2.A	-1.82	0.01	1.26	1.20	0.61	0.66

Table 10
Scale scores by domain and assessment period

		Fall	Winter	Spring	Growth
Social Emotional	Mean	417.21	473.13	529.11	112.47
	SD	59.61	70.91	80.19	77.43
	25th percentile	377	435	487	67
	Median	422	472	537	118
	75th percentile	450	519	586	158
	SEM	15	16	18	
Physical	Mean	586.47	669.79	738.75	154.63
	SD	90.78	93.50	102.71	100.22
	25th percentile	511	614	702	93
	Median	595	677	767	167
	75th percentile	634	727	806	216
	SEM	28	31	30	
Language	Mean	438.15	510.54	576.61	140.50
	SD	90.24	95.91	109.17	89.14
	25th percentile	382	455	526	94
	Median	455	526	601	145
	75th percentile	480	574	631	196
	SEM	20	19	20	
Cognitive	Mean	503.45	584.38	657.90	156.69
	SD	80.31	92.91	106.92	93.98
	25th percentile	448	535	600	104
	Median	515	600	672	165
	75th percentile	557	655	736	215
	SEM	18	20	18	
Literacy	Mean	670.69	754.63	783.88	125.73
	SD	79.11	60.47	55.46	62.83
	25th percentile	626	731	763	85
	Median	682	765	793	121
	75th percentile	720	791	818	160
	SEM	17	14	13	
Mathematics	Mean	548.66	621.33	648.78	131.58
	SD	81.96	72.26	87.11	71.86
	25th percentile	510	591	591	92
	Median	559	629	669	125
	75th percentile	607	669	714	171
	SEM	16	16	15	

Table 11
Percetnages below, meeting, or exceeding Widely Held Epxectations

		Fall	Winter	Spring
Social Emotional	Below	61.34	23.39	9.90
	Meet	37.96	73.61	75.02
	Exceed	0.70	3.00	15.08
	Meet or Exceed	38.66	76.61	90.10
Physical	Below	55.54	21.79	9.01
	Meet	43.73	75.04	72.94
	Exceed	0.73	3.17	18.04
	Meet or Exceed	44.46	78.21	90.99
Language	Below	75.24	36.54	18.01
	Meet	23.64	58.55	59.55
	Exceed	1.12	4.91	22.43
	Meet or Exceed	24.76	63.46	81.99
Cognitive	Below	74.70	32.00	13.76
	Meet	24.99	67.03	78.12
	Exceed	0.31	0.97	8.11
	Meet or Exceed	25.30	68.00	86.24
Literacy	Below	45.60	7.75	2.49
	Meet	52.53	75.02	47.86
	Exceed	1.87	17.24	49.65
	Meet or Exceed	54.40	92.25	97.51
Mathematics	Below	33.89	11.99	3.12
	Meet	55.62	43.30	30.14
	Exceed	10.49	44.72	66.75
	Meet or Exceed	66.11	88.01	96.88

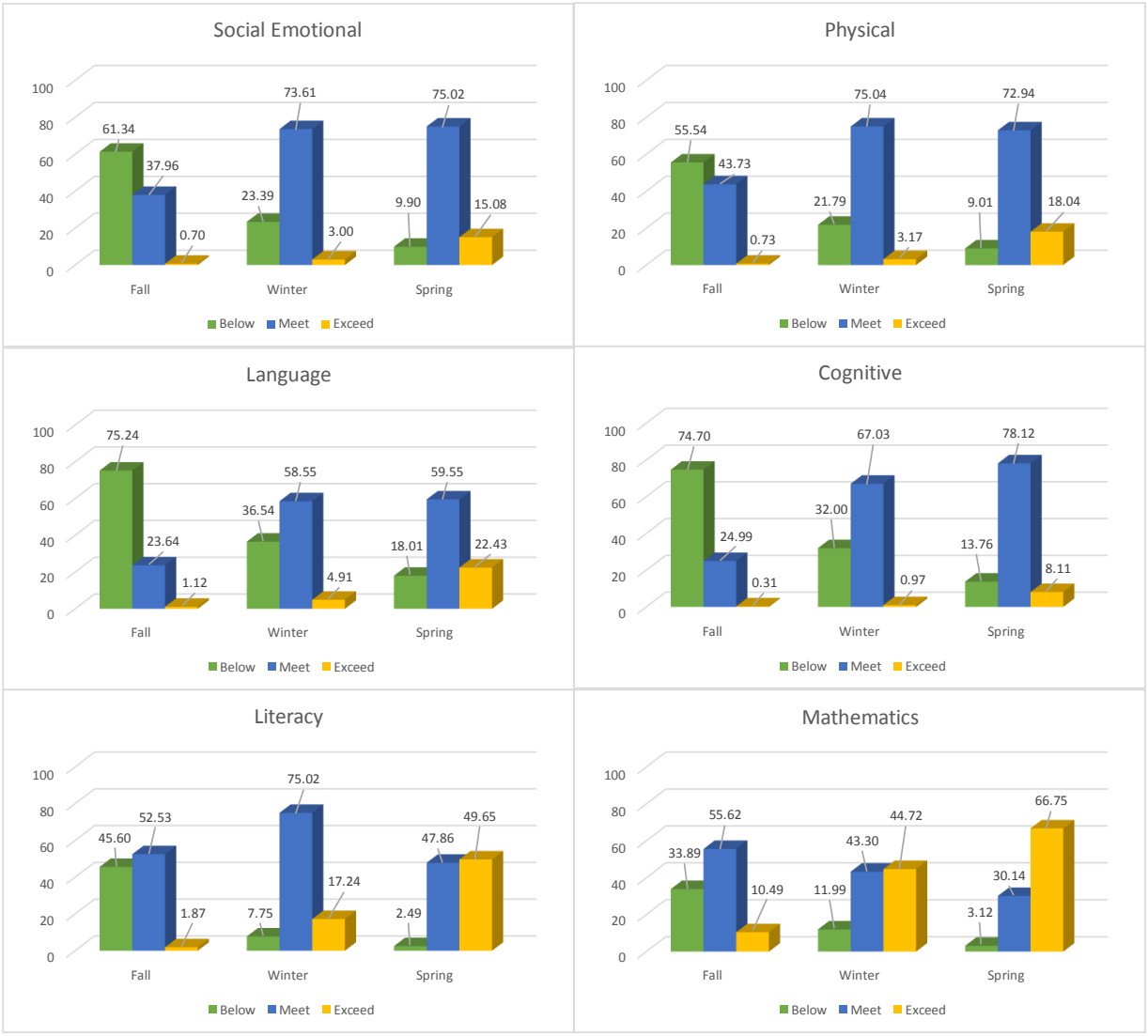

Figure 1. Widely held expectations by domain of development and assessment period.

Table 12
Correlations between scale scores across assessment periods

		Fall	Winter	Social Emotional	Physical	Language	Cognitive	Literacy	Mathematics
Social Emotional	Fall				.695	.701	.750	.607	.534
	Winter	.639			.666	.681	.743	.567	.543
	Spring	.463	.740		.649	.685	.740	.604	.485
Physical	Fall			.695		.643	.656	.532	.483
	Winter	.602		.666		.624	.659	.491	.500
	Spring	.424	.721	.649		.674	.747	.532	.568
Language	Fall			.701	.643		.794	.728	.649
	Winter	.721		.681	.624		.776	.674	.640
	Spring	.601	.771	.685	.674		.830	.698	.641
Cognitive	Fall			.750	.656	.794		.755	.679
	Winter	.641		.743	.659	.776		.676	.674
	Spring	.495	.783	.740	.747	.830		.697	.662
Literacy	Fall			.607	.532	.728	.755		.748
	Winter	.704		.567	.491	.674	.676		.758
	Spring	.622	.842	.604	.532	.698	.697		.772
Mathematics	Fall			.534	.483	.649	.679	.748	
	Winter	.687		.543	.500	.640	.674	.758	
	Spring	.573	.787	.485	.568	.641	.662	.772	