Technical Manual for the Teaching Strategies *GOLD*® Assessment (2nd Edition)

Birth Through Third Grade

October 2020

Prepared for Teaching Strategies
by Richard G. Lambert, Ph.D.
Center for Educational Measurement and Evaluation
University of North Carolina Charlotte

Suggested citation:
Lambert, R. (2020). *Technical manual* for the Teaching Strategies GOLD® *assessment (second edition): Birth through third grade*. Center for Educational Measurement and Evaluation, University of North Carolina Charlotte.

Technical Manual for the Teaching Strategies *GOLD*® Assessment (2nd Edition)

Birth Through Third Grade

Richard G. Lambert, Ph.D.
Center for Educational Measurement and Evaluation UNC Charlotte
October, 2020

UNC CHARLOTTE
CATO COLLEGE *of* EDUCATION

CEME
The Center for Educational Measurement and Evaluation

Table of Contents

Purpose of the New Technical Manual

Teaching Strategies GOLD® is a formative assessment intended to assess the whole child from birth through third grade. It enables teachers to collect documentation on an ongoing basis to identify the best placements for individual children across a series of developmental progressions. In contrast to direct assessments, teachers collect evidence during regular activities in natural classroom contexts. Assessment, thus, unfolds as teachers compile portfolios of evidence for each child, reflect upon and analyze the evidence, make preliminary ratings on a rolling basis, and finalize ratings at specified points during the year. Teachers and administrators may then use this information to inform instruction and to facilitate communication with parents and other stakeholders. The primary intent of *GOLD®* is to help teachers observe and understand child progress, plan instruction, and support child growth and development. Moreover, unlike direct assessment, the process gives young children agency by directly involving them in meaningful interactions that signal developmental progress.

It is important to provide teachers with formative assessments that yield reliable, valid, and culturally sensitive information about children. 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. Hence, the measurement properties of any assessment should be rigorously examined as long as it is in use, and this examination must extend to all subgroups of children and all conditions of use. Previous studies of *GOLD®* have demonstrated its reliability and validity overall and for multiple subgroups of interest. For a thorough review, see Lambert, Kim, & Burts, 2013; Lambert, Kim, & Burts, 2014; Lambert, Kim, & Burts, 2015; and the third technical manual for the original edition (birth through kindergarten) of *GOLD®*.

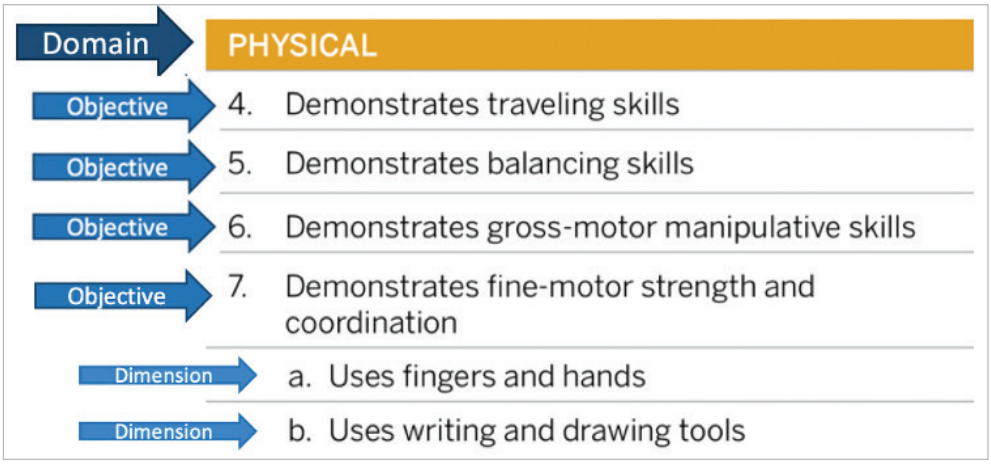
This is the second technical manual for the birth through third grade (B–3) edition of *GOLD®*; the first was released in 2017 alongside the B–3 edition. While nothing about the assessment has changed, the aims of the current manual are to:

1. Reaffirm the reliability and validity of *GOLD®* on a wider range of data collected by more experienced users. With these improvements to the sample, the current manual allows for deeper understanding of the assessment's functionality.
2. Provide updated norms, including an expansion into first grade.
3. Outline the improved approach to imputing missing data.

The Design of *GOLD*®

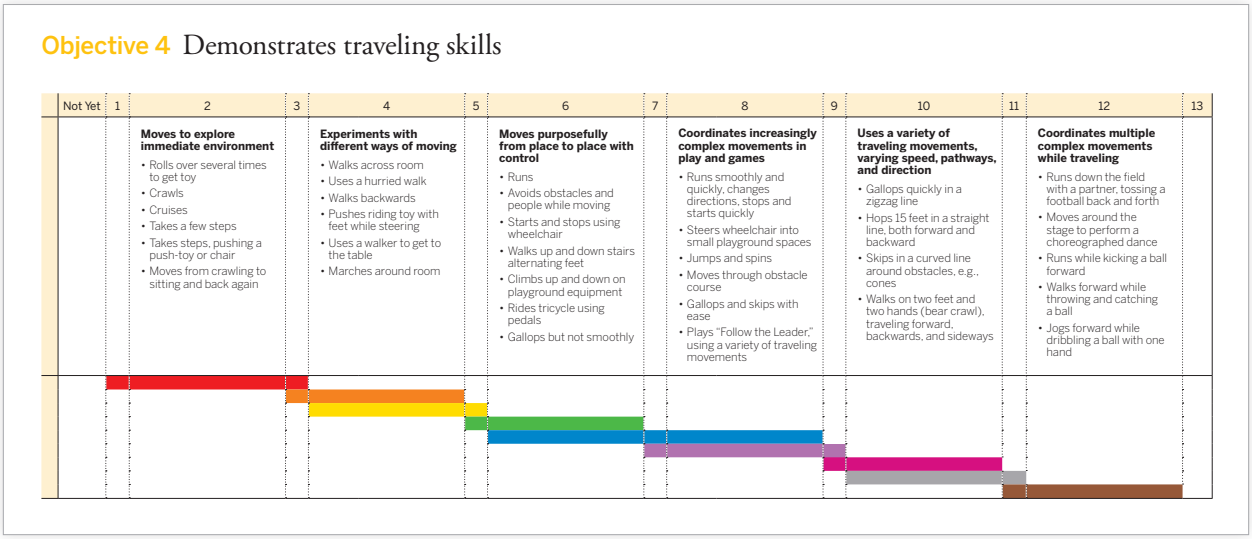
Taking a whole-child approach, *GOLD*® assesses children's development and learning across four **developmental domains** (social–emotional, physical, language, cognitive) and five **content domains** (literacy, mathematics, science and technology, social studies, and the arts). It also includes a tenth domain, English language acquisition, for use with dual-language learners (DLLs). Each domain comprises a set of **objectives** designed to guide teachers through the assessment process. Many objectives are further broken down into one or more **dimensions**. *GOLD*® has thirty-eight objectives in total, collectively termed the Objectives for Development and Learning (ODL). Figure 1 illustrates the organization of the ODL for the physical domain as an example.

Figure 1. ODL for the Physical Domain



In addition, each dimension under the first six domains (social–emotional, physical, language, cognitive, literacy, and mathematics) has a progression associated with it. Every **progression** represents a continuum that enables teachers to relate observable behavior to expectations for a child's age/grade. Progressions help teachers and other stakeholders understand how children are performing relative to developmentally appropriate expectations. Figure 2 illustrates an example of the progression for Objective 4 under the physical domain.

Figure 2. Progression Example: Objective 4 Under the Physical Domain



As exemplified above, *GOLD*® uses colored bands to represent the associated developmental and learning expectations for a given year of life or academic program year. Each band corresponds to the range of widely held expectations (WHE) for a particular age/grade. Knowledge, skills, and abilities that fall within a child's respective colored band are meeting widely held expectations for that objective or dimension. Those that fall to the left of the colored band are below expectations, and those that fall to the right are exceeding expectations. Colored bands that fall completely under the "Not Yet" level indicate that it is not yet developmentally appropriate to expect children to demonstrate the skill(s). WHE were developed by panels of experts in child development based on the latest developmental theory and research. This manual focuses on the six domains that have progressions for which WHE have been developed.

Figure 3. GOLD® Colored Bands

- Birth to 1 year
- 1 to 2 years
- 2 to 3 years
- Preschool 3 class
- PreK 4 class
- Kindergarten
- First Grade
- Second Grade
- Third Grade

The Purpose of *GOLD*®

Users must understand the central purpose of any assessment to ensure that they actually assess the intended outcomes. Therefore, it is valuable to delineate appropriate and inappropriate uses. *GOLD*® has been designed and externally validated for use as a formative, developmental, authentic, and criterion-referenced assessment. Its primary purpose is to provide teachers with instructionally relevant information about the children they teach. Figure 4 contrasts optimal and ineffective uses of *GOLD*®, and the sections below it describe the most meaningful applications of the information it generates.

***Figure 4*. Optimal and Ineffective Uses of *GOLD*®**

Optimal Uses of *GOLD*®	
***GOLD*® is a Formative Assessment**	• Formative assessment focuses on the learning process and is used to support learning while learning is taking place. • Formally, formative assessment is "...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*® is a Developmental Assessment**	• *GOLD*® includes progressions of growth, development, and learning that describe a sequence of stages and behavioral anchors that children are generally expected to demonstrate at a given age/grade. • It is designed to help teachers assess the whole child.
***GOLD*® is an Authentic Assessment**	• Authentic assessments help teachers observe the progress children make through a process that emerges naturally. • Evidence of development and learning is gathered during everyday instructional activities and routines.
***GOLD*® is a Criterion-Referenced Assessment**	• *GOLD*® measures developmental progress and learning relative to a fixed set of standards, i.e., the ODL. • One child's performance on *GOLD*® does not depend on all others'.

GOLD® Is a Formative Assessment

GOLD® is a valuable resource for teachers getting to know children at the beginning of the school year. It can help them understand the strengths that a child brings to the classroom as well as areas where a child needs support. It is also useful at any point in the academic year, allowing teachers to understand the current status of the growth, learning, and development of the children in their classrooms. Finally, it can help teachers identify children's interests, plan classroom activities, select and rotate classroom materials, and individualize and differentiate instruction.

When teachers communicate with parents and other guardians, *GOLD*® can help them do so in terms that can be easily accessed and understood. Teachers can link specific examples of what a child knows and can do with placements on the developmental progressions. This process can help facilitate rich conversations about the child's development within the classroom, family, and cultural context. Teachers can even solicit evidence of child progress and development from parents and other caregivers to inform the assessment process and ensure that progress is measured accurately. This process can help teachers partner with parents to support the growth and development of each child.

Other advantages of *GOLD*® include helping teachers collaborate with other educational professionals who serve the same children, identifying areas of need for professional development and other training, empowering mentors to better support teachers with planning instructional activities, and enhancing observational skills and awareness of how children learn and function in the classroom. In summary, *GOLD*® is especially effective at supporting data-driven decisions around individualizing support for children.

GOLD® Is a Developmental Assessment

GOLD® has been aligned to the Head Start Early Learning Outcomes Framework, the Common Core Standards, and state early learning standards. Most standards outline specific skills and abilities that children are expected to obtain by the end of a particular grade or age level. *GOLD*® expresses these standards not just in terms of the end point, but also 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 at least one developmental progression. Each progression details the steps children are expected to follow on their way to mastery of new skills and abilities. In this way, the assessment helps teachers connect and understand the relationship between learning standards, curricular

goals, and instructional objectives in terms of evidence that can be observed within everyday classroom contexts. Most importantly, it helps teachers individualize the implementation of scaffolds each child needs to reach the next step of their developmental trajectory.

GOLD® Is an Authentic Assessment

Authentic assessments help teachers observe the progress children are making through a process of gathering evidence of learning as it emerges naturally throughout the course of typical classroom activities. Evidence documented by teachers is then used to assess where children are on a progression and then to support further development along the progression. 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). The authentic process of formative assessment has often been described as a continuous cycle of activities that are part of everyday instruction (Heritage, 2013). This cycle is often outlined in phases, as illustrated in Figure 5.

Figure 5. The Authentic Assessment Cycle

GOLD® Is a Criterion-Referenced Assessment

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. Teachers can use the widely held expectations for each colored band to understand which behaviors and skills children can generally be expected to demonstrate in the classroom. These expectations are not grounded 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 empirical research. While criterion-referenced assessments are not norm-referenced assessments, the norms that are available toward the end of this manual are designed to provide teachers with an additional resource. Ultimately, the norms are for those who are interested in a broad and comprehensive picture of how a child is growing and developing relative to the developmental progress of other children across the nation.

Using the Scores Provided by the *GOLD*® Assessment System

There are several types of scores for the placements that teachers make on the developmental progressions. These include **raw scores, widely held expectations scores, scaled scores,** and **national norm scores**. Each of these scores will be described in this section of the manual along with guidelines for their appropriate use.

Raw Scores

GOLD® raw scores are calculated from the sum of all the finalized ratings a teacher made on the progressions under each domain. Raw scores, hence, represent a child's current knowledge, skills, and abilities related to a particular domain of development. At the end of each checkpoint a child may have up to six different raw scores. As the progressions do not have a consistent number of steps, and one step on any given progression does not represent the same amount of growth and development, these total raw scores need to be transformed into scaled scores to measure the amount a child has grown.

Widely Held Expectations Scores

Widely held expectations reflect the expected developmental trajectories for children on each dimension. Each colored band corresponds to the knowledge, skills, and abilities children are expected to demonstrate over a given year of life or from the beginning to the end of a program year. As already mentioned, if a child's rating

falls within the colored band that corresponds to her age/grade, that child's skills are meeting WHE for the dimension. If her rating falls to the left of the colored band, her skills are below expectations. Likewise, if it falls to the right, her skills are exceeding expectations. Generating *GOLD®* reports that display WHE in *MyTeachingStrategies®* enables teachers to compare data for specific children or groups of children to determine if their knowledge, skills, and abilities are below, meeting, or exceeding expectations. *MyTeachingStrategies®* offers teachers the ability to make these comparisons at the level of individual dimensions and at the domain level.

To find the range of WHE for a given domain, sum the numeric values on each progression that correspond to the bottom level of each colored band; this total assigns the lower bound of the domain's WHE. Then, sum the numeric values on each progression that correspond to the uppermost level of each colored band. Similarly, this total assigns the upper bound of the domain's WHE. For example, for Pre-K 4 (the blue colored band), WHE for the physical dimensions range from 6–8 for Objective 4, from 6–8 for Objective 5, from 6–8 for Objective 6, from 6–8 for dimension 7.A, and from 5–7 for dimension 7.B. Therefore, the range of widely held expectations for the blue band is (6 + 6 + 6 + 6 + 5) to (8 + 8 + 8 + 8 + 7), or 29–39. If the raw score for a Pre-K 4 child's physical abilities is 33, which falls within the range of 29–39, her performance meets WHE for the physical domain. Tables 16 through 21 present the proportion of children who are below, meeting, and exceeding WHE in the fall, winter, and spring based on a nationally representative sample for each age/grade.

Scaled Scores

It is important to recognize that raw scores are intended for use only in generating domain-level WHE scores. Raw scores alone are not useful for tracking growth and development over time for the following reasons. First, the progressions within a given domain of development do not include the same number of steps and, therefore, do not contribute the same number of potential points to the total raw score. Second, the raw scores for each progression represent an ordinal level of measurement. This means that a score of 2 represents a higher level of development than a score of 1, and a lower level of development than a score of 3, etc. However, these scores cannot be interpreted such that an increase of one point represents an equal amount or "quantity" of development. In order to facilitate the interpretation of a child's growth over time in terms of equal intervals, or score points, raw scores must be converted to scaled scores.

GOLD® has been designed and validated to be used primarily as a criterion-referenced, authentic, formative assessment. Nevertheless, two types of scores are provided to users for the purpose of specific norm-referenced interpretations: scaled scores and national norm scores. Developmental scaled scores are provided for the purposes of converting raw scores from each domain of development to a common scale. These scaled scores eliminate the impression that one domain of development is more important than another. Scaled scores are primarily provided for the purpose of tracking a child's growth and development across different age/grade levels, i.e., they allow teachers to compare the amount of growth a child made over different time periods. For example, scaled scores can tell a teacher whether a child grew more/less over the spring checkpoint period than she did over the fall checkpoint period or whether she grew more/less throughout one school year versus another. Scaled scores also allow for comparisons between children, even if they are at different levels of development. Note that comparisons may only be made between scaled scores from the same domain. The *GOLD*® uniform scale ranges from 0 to 1,000.

National Norm Scores

GOLD® scaled scores are also used to create national norm scores to facilitate interpretations about how children of a particular age range tend to be rated by other teachers across the nation. National norm scores are based on quartiles, or fourths, of the national distribution of scaled scores for each colored band for a given domain. Children with scaled scores in the first quartile (lowest 25%) are considered below the national norm, children with scores in the second and third quartiles (the middle 50%) are considered within the national norm, and children with scores in the fourth quartile (highest 25%) are considered above the national norm. These scores can be a valuable resource for teachers interested in understanding how the children they serve have been placed on the progressions relative to how teachers across the nation place their own children. To that end, updated national norms are provided in Tables 10 through 15.

Guidelines for Aggregated Reporting

The most important considerations with regard to reporting assessment scores are 1) the purpose, 2) the specific inferences that are to be made from the scores, and 3) whether these align with the nature of the assessment in question. With respect to *GOLD*®, the assessment was designed—and has been extensively validated—as an

authentic, developmental, formative assessment. Its primary purpose is to provide actionable information for teachers as they plan instruction and support the growth and development of young children.

It is possible to use aggregated *GOLD*® data in a responsible way. However, if users are considering aggregated reporting, various threats to the validity of the inferences made from the scores need to be given careful consideration. The teachers producing such data must have received all appropriate training, including ongoing coaching. It is also preferred that they have achieved interrater reliability certification. Reliable assessors are just as likely to be strict as they are to be lenient in their placements of children on the developmental progressions. Still, some teachers' placements may be consistently strict or lenient. Nevertheless, with adequate training, any measurement error caused by rater effects should wash out in the aggregate. It is essential that this assumption holds to ensure responsible use of aggregated scores on *GOLD*®.

About Missing Data

GOLD® requires teachers, as part of the ongoing assessment process, to collect evidence of each child's developmental progress during the normal course of classroom activities. For some assessment periods, teachers may not have the opportunity to collect sufficient evidence for a child to support a finalized placement on a particular progression. Some children join a classroom later in the assessment period; others miss classroom activities that generate evidence pertaining to a given progression. Since placements (or "ratings") tend to be highly correlated with each other within a given domain of development, imputation of reasonable amounts of missing data can allow a total score to be obtained. Note that any imputation is domain-specific, i.e., when an item is missing a rating, only items from within the same domain are used to impute its value.

Teaching Strategies has developed an imputation method that takes into account not only the varying number of steps on the progressions but also the varying locations and widths of the colored bands. Termed tripartite weighting, this method involves using each rating's relative placement with respect to the total number of steps in its corresponding region of the progression to impute missing ratings. Progressions are divided into three regions relative to each colored band: below the band, within the band, and above the band. Each region may have a different number of steps, dependent on the difficulty of the item and the developmental pathway of the construct it measures. If a child receives a rating of 3 that happens to fall in the region

below a colored band that begins at 6, for example, then the rating's relative placement on the region below the colored band is 0.6. Likewise, if a child receives a 5 and the rating falls within a colored band that spans scores 4 through 8, then the rating's relative placement on the region within the colored band is 0.25. Any missing rating is estimated once per relative placement of all the ratings that are present. The average of these estimations yields the imputed rating for a given item. For instance, if a child is missing one out of 10 ratings for a given domain, then that item's imputed rating will result from the average of 9 estimations; if a child is missing two out of 10 ratings, then both items' imputed ratings will result from separate averages of 8 estimations. Total raw scores may be calculated by summing the original and imputed ratings and rounding down the result. *GOLD*® scaled scores may then be calculated as normal from the resulting raw scores.

The Current Study

This study examined *GOLD*® assessment records from the 2018–2019 academic year. Over 1.5 million American children were assessed at some point during the year, providing a broad pool from which to draw a nationally representative sample. Study methods involve the use of Rasch measurement models to examine the properties of the information provided by each developmental progression. Specifically, results are first presented for dimensionality to confirm the appropriateness of the model, followed by analysis of rating scale validity, item difficulty measures, and reliability measures.

National Sample

The sample for this study was drawn from the more limited population of 842,336 children who were all assessed on *GOLD*® during fall, winter, and spring of 2018–2019. Teachers in programs that assess children in all three semesters are more likely to understand the process of collecting valid evidence as well as how to translate that evidence into meaningful placements on the developmental progressions. In other words, their ratings are likely more reliable than ratings assigned by teachers who assess children only once or twice per year. Nevertheless, a few qualifiers about the sample should be mentioned. Not all of the programs have adopted the *GOLD*® interrater reliability certification process, and others are in various stages of requiring and/or achieving this certification for all of their teachers. In addition, teachers working with children who are in kindergarten through third grade would, of course, use the curriculum that is outlined for them by their local education

agency. Therefore, they do not use Teaching Strategies' *The Creative Curriculum®*. Conversely, many of the programs serving younger children were also users of *The Creative Curriculum®*.

To form the whole sample, a stratified random sample of 5,000 children was taken from each birth to kindergarten age/grade group (colored band). The strata were formed using the U.S. Census Bureau's data for race/ethnicity in an effort to represent each subgroup in proportion with the U.S. population of children. Incidentally, this process mitigates any clustering of children within their rater or teacher, as whole classrooms of children were not selected. The *GOLD®* assessment has not been adopted for use in first, second, and third grades at the same rate as it has for the birth to kindergarten years. Therefore, all available data from children in first, second, and third grade who have fall, winter, and spring assessment records was used, regardless of school year. This process, as shown in Table 1, resulted in a total norm sample of 32,063 children.

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 six race/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. As shown in Table 2, the norm sample was roughly balanced by gender (boys = 48.52%, girls = 51.48%). Children with an IFSP or IEP comprised 7.24% of the sample. A total of 33.19% of the norm sample qualified for the National School Lunch Program (free or reduced-price lunch) as reported by their teacher. This figure is likely an underestimate as all teachers may not accurately report this information. As is, this number reflects an under-representation of these children, as the national figure is over 50%. The primary language spoken in the home was distributed as follows: English (78.57%), Spanish (14.47%), and other languages (6.96%). The race/ethnicity of the children in the sample was as follows, shown here with the 2018 national census estimates for U.S. children in parentheses: a) White – 49.69% (49.39%), b) African American – 13.89% (13.76%), c) Native American/Pacific Islander – 1.22% (1.04%), d) Asian – 4.13% (5.04%), e) Multiracial/other – 4.64% (4.70%), and f) Hispanic – 26.43% (26.07%). These values indicate that the sample was approximately nationally representative for all race/ethnicity groups of American children.

Furthermore, the sample of children from birth to kindergarten received educational services in programs that were located in the all 50 U.S. states and the District of Columbia, Puerto Rico, and in limited locations around the world that serve the families of U.S. military personnel. The children for the first, second, and third grade sample were from five states: Alabama, Colorado, Georgia, Hawaii, and New Jersey. Therefore, the sample includes a rich geographic diversity from all regions of the U.S. and includes all types of settings where young children are served (Head Start, private prekindergarten, publicly funded preschool programs, and public schools).

Analyses Related to the Construction of Scaled Scores

Rasch models were used to create ability estimates for each child on each domain and to examine the measurement properties of the information provided by each developmental progression. *GOLD*® assessment data were analyzed using the Rasch 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 model (RSM; Bond & Fox, 2001) and the PCM are the two most widely used Rasch models for polytomous response data. The PCM, rather than the RSM, was chosen because the items do not share the same rating scales (i.e., the number of steps is different across items). Specifically, 11 *GOLD*® items are on a 0–9 scale, six items are on a 0–11 scale, 17 items are on a 0–13 scale, 25 items are on a 0–15 scale, and one item is on a 0–19 scale. For each item, the 0 category represents "Not Yet," and the maximum category represents abilities beyond the most developmentally advanced behavioral anchor.

Dimensionality

Rasch modeling assumes unidimensionality, meaning that the items under a given domain measure only one latent construct. The unidimensionality of each scale was evaluated via mean-square (MNSQ) fit statistics and principal components analysis of Rasch residuals (PCAR). MNSQ fit values between 0.6 and 1.4 are considered reasonable for rating scale items (Bond & Fox, 2007). MNSQ values still 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. In other interpretations, items with mean square values of between 1.4 and 2.0 can be considered potentially unproductive for the development and construction of new 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 fit to the underlying construct and the possibility of secondary dimensions. 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 less than 3 and accounts for less than approximately 5% of the unexplained variance, unidimensionality is considered plausible (Linacre, 2012). The following sections present the results of the MNSQ fit and PCAR analyses; note that indices were evaluated for the fall assessment checkpoint.

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 (91.9%) with an eigenvalue of 101.4. The first contrast (the largest secondary dimension) had an eigenvalue of 1.9 and accounted for 1.8% of the unexplained variance. All fit statistics for all of the social–emotional items were within acceptable limits. The infit MNSQ values ranged from 0.85 to 1.18. The outfit MNSQ values ranged from 0.83 to 1.20. The item total score correlations, with each item sequentially excluded from the total score, ranged from .93 to .95.

Physical Scale (5 items)

The PCAR showed that for the physical scale, the Rasch dimension explained the majority of the variance in the data (92.9%) with an eigenvalue of 65.0. The first contrast had an eigenvalue of 1.6 and accounted for 2.2% unexplained variance. All fit statistics for all of the physical items were within acceptable limits with one exception. The infit MNSQ values ranged from 0.84 to 1.22. The outfit MNSQ values ranged from 0.81 to 2.22. Item 7.B had an outfit statistic just outside the optimal range (2.22), suggesting the presence of outliers or unusual response patterns. The item total score correlations, with each item sequentially excluded from the total score, ranged from .94 to .96.

Language Scale (8 items)

The PCAR showed that for the language scale, the Rasch dimension explained the majority of the variance in the data (93.4%) with an eigenvalue of 113.8. The first contrast had an eigenvalue of 1.5 and accounted for 1.2% of the unexplained

variance. All fit statistics for all of the language items were within acceptable limits. The infit MNSQ values ranged from 0.78 to 1.15. The outfit MNSQ values ranged from 0.79 to 1.17. The item total score correlations, with each item sequentially excluded from the total score, ranged from .94 to .96.

Cognitive Scale (10 items)

The PCAR showed that for the cognitive scale, the Rasch dimension explained the majority of the variance in the data (93.3%) with an eigenvalue of 140.4. The first contrast had an eigenvalue of 1.9 and accounted for less than 1.2% of the unexplained variance. All fit statistics for all of the cognitive items were within acceptable limits. The infit MNSQ values ranged from 0.79 to 1.07. The outfit MNSQ values ranged from 0.79 to 1.12. The item total score correlations, with each item sequentially excluded from the total score, ranged from .93 to .96.

Literacy Scale (16 items)

The PCAR showed that for the literacy scale, the Rasch dimension explained the majority of the variance in the data (92.2%) with an eigenvalue of 189.8. The first contrast had an eigenvalue of 2.0 and accounted for 1.0% of the unexplained variance. All fit statistics for all but two of the literacy items were within acceptable limits. The infit MNSQ values ranged from 0.70 to 1.47. The outfit MNSQ values ranged from 0.68 to 4.78. Items 18.D (4.78) and 15.C (2.03) had outfit MNSQ values just above 1.4, suggesting the presence of outliers or unusual response patterns. The item total score correlations, with each item sequentially excluded from the total score, ranged from .55 to .94.

Mathematics Scale

The PCAR showed that for the mathematics scale, the Rasch dimension explained the majority of the variance in the data (90.3%) with an eigenvalue of 112.0. The first contrast had an eigenvalue of 2.4 and accounted for 2.0% of the unexplained variance. All fit statistics for the mathematics items were largely within acceptable limits. The infit MNSQ values ranged from 0.78 to 1.37. The outfit MNSQ values ranged from 0.71 to 2.19. Two items had outfit MNSQ values above 1.4: 20.D (2.19) and 22.B (1.83). These values suggest the presence of outliers or unusual response patterns. The item total score correlations, with each item sequentially excluded from the total score, ranged from .57 to .94.

In summary, with the few exceptions noted above, the model fit statistics generally suggest that the data does in fact fit the Rasch PCM very well. The results indicate, namely, that the data satisfy the unidimensionality assumption of the Rasch model.

Rating Scale Category Effectiveness

The rating scale categories were also examined to provide insight into whether teachers use the instrument in the manner in which it was intended. Rating scale category effectiveness is one way to measure the validity of the developmental progressions. The averages of the ability estimates for each domain (based on the total item scores) were examined for each group of children placed at a particular rating on a given developmental progression. The averages should advance monotonically with rating scale category values (Bond & Fox, 2007). Andrich thresholds (also called step calibrations) from the PCM were also analyzed (1978). Thresholds parameterize, relative to item difficulty, the points at which a rater is equally likely to select one rating or its neighboring rating along the latent variable underlying the progression (Bond & Fox, 2007). Thresholds should also increase monotonically along the rating scale categories. This means that the probability that a child is placed on the next step on the developmental progression should always increase as the overall ability level of the child increases.

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. There were no disordered category thresholds. The observed sample averages for each response category generally advanced as expected across the rating scales with no disordered steps. For 7 of the 9 items there were no disordered average placements. However, for two of the progressions (1.A and 2.A) there were disordered averages at the top of the scale. This means that the average total score for children placed at the top of the scale was below the average of those placed at the next to the highest step on the progression. The highest step on all *GOLD*® progressions does not have behavioral anchors. Rather, it is to be used when a child's behavior exceeds expectations consistent with the behavioral anchors for the penultimate rating and is considered by the teacher to have advanced in their development beyond the scope of the progression.

It is important to note that these statistics are sample-specific. The sample used for this study contained a relatively small number of second-grade children and an even smaller number of third-grade children, and those children who were included were from a narrow range of classrooms. Therefore, these results may be an artifact of the limited sample of older children. It will be important to monitor these finding in future studies, especially after attaining larger and more diverse samples of second- and third-grade children. Similar findings of disorder at the top of the scale appear across the other domains and must be considered in light of these sample limitations.

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. There were no disordered category thresholds for three of the five progressions. Two progressions, 5 and 6, had disordered thresholds at the top of the progression. The observed sample averages for each response category advanced as expected across the rating scales with no disordered averages for three of the progressions. There were, however, disordered averages at the top of the scale for three of the progressions (6, 7.A, and 7.B).

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. There were no disordered category thresholds for three of the eight progressions. Five of the progressions (8.A, 9.A, 9.B, 9.C, and 10.B) had disordered thresholds at the top of the scale. The observed sample averages for each response category generally advanced as expected across the rating scales with no disordered steps for two of the eight progressions. However, for six progressions (8.A, 9.A, 9.B, 9.D, 10.A, and 10.B), there were disordered averages at the top of the scale.

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. There were no disordered category thresholds for eight of the ten progressions. Two progressions (14.A and 14.B) had disordered thresholds at the top of the scale. The observed sample averages for each response category advanced as expected across the rating scales with no disordered steps for six of the ten progressions. However, four progressions (11.A, 11.E, 12.B, and 14.B), there were disordered averages at the top of the scale.

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. For the literacy scale, there were issues with the category thresholds for the majority of items. For 10 of the 16 developmental progressions, there were disordered thresholds at the top of the scale. The observed sample averages for each response category generally advanced as expected across the rating scales. However, four of the 16 developmental progressions (15.C, 17.A, 18.A, and 18.D) had disordered averages at the top of the scale.

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. For the mathematics scale, there were no issues with the category thresholds for 11 of the 12 progressions. One progression (20.D) had disordered thresholds at the top of the scale. There were issues with disordered sample averages for the response categories at the top of the rating scales for 6 of the 12 developmental progressions (20.C, 20.D, 20.E, 20.F, 21.B, and 23).

These results suggest potential issues with assessing older children on some of the progressions. As already noted, it may well be the case that these results are purely an artifact of the limited sample of older children. If they were to hold in a broader sample, however, they would indicate several avenues for additional investigation. The training and implementation processes for the teachers of older children may need to be evaluated and monitored. The *GOLD*® approach is likely novel to many teachers in the early elementary grades. It is certainly plausible that second- and third-grade teachers need more support as they collect, interpret, and analyze their documentation. Alternatively, the utility and appropriateness of the behavioral anchors at the top of many of the progressions may require more in-depth examination. Ultimately, additional analysis of larger and more diverse samples of children in the early elementary grades is indicated.

Item Difficulty Measures

The Rasch measurement model estimates the item difficulty level for each of the progressions within a given domain of development. These estimated difficulty levels are translated into item locations on the same metric as the child ability estimates.

Individual progressions can then be compared according to their relative difficulty levels and can be evaluated and interpreted as a developmental pathway from the least difficult item to the most difficult item. It is expected that most children will develop the skills and abilities associated with progressions located lower on the difficulty metric earlier than they will develop those associated with progressions located higher on the metric. For all six domains, the item location hierarchy appeared to be generally consistent with the expected developmental trajectory for typically developing children. Hierarchies were also consistent with findings from previous studies. 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 estimates were developed for the winter assessment period.

For the physical scale, the item pertaining to a child's ability to use their fingers and hands (7.A) was estimated as the easiest item (-1.03). The item pertaining to a child's ability to use writing and drawing tools (7.B) was estimated as the most difficult item (1.27). The range of item difficulties (-1.03 to 1.27) and item rating scale threshold locations (-16.27 to 11.32) 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 follow directions (8.B) was estimated as the easiest item (-2.80). The item pertaining to a child's ability to tell about another time or place (9.D) was estimated as the most difficult item (1.06). The range of item difficulties (-2.80 to 1.06) and item rating scale threshold locations (-15.97 to 11.14) 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 persist within classroom activities (11.B) was estimated as the easiest item (-1.60). The item pertaining to a child's ability to make connections between different experiences (12.B) was estimated as the most difficult item (0.97). The range of item difficulties (-1.60 to 0.97) and item rating scale threshold locations (-17.73 to 13.83) was considered wide enough for reasonable separation of children according to underlying ability.

For the literacy scale, the item pertaining to a child's ability to write their name (19.A) was estimated as the easiest item (-2.88). The item pertaining to a child's ability to use context clues to read and comprehend texts (18.D) was estimated as

the most difficult item (2.99). The range of item difficulties (-2.88 to 2.99) and item rating scale threshold locations (-14.34 to 9.27) 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 understand spatial relationships (21.A) was estimated as the easiest item (-4.11). The item pertaining to a child's ability to apply properties of mathematical operations and relationships (20.E) was estimated as the most difficult item (2.37). The range of item difficulties (-4.11 to 2.37) and item rating scale threshold locations (-9.30 to 17.63) 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 aligns with expectations from developmental theory. In addition, the range of difficulties for each scale is the widest that has been observed across various validation studies. This finding suggests that teachers are able to separate children according to their analysis of the evidence they collected.

Reliability

Reliability refers to how consistently a particular construct is measured by an assessment. This consistency can be measured by examining the internal consistency across items within a scale, the level of agreement between raters, or the extent to which scores for a child replicate across different assessments, testing situations, and/or time points. Reliability is considered an important psychometric characteristic of the information an assessment provides and is a necessary precondition for the validity of the information provided by any assessment.

Reliability was evaluated using the following Rasch indices: 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. The person separation index indicates how well the instrument can differentiate persons on each of the constructs. The item separation index indicates the spread or separation of items along the measurement constructs. Reliability separation indices greater than 2 are considered adequate, and indices 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 multiple 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-retest 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 scaled scores. Across all domains of development, 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 indices for all domains of development were very high and are therefore not included in the table. Specifically, for the social–emotional scaled scores, all of the item separation indices, both sample-based and model-based, were greater than 150. For the physical scaled scores, all of the item separation indices, both sample-based and model-based, were greater than 100. For the language scaled scores, all of the item separation indices, both sample-based and model-based, were greater than 120. For the cognitive scaled scores, all of the item separation indices, both sample-based and model-based, were greater than 80. For the literacy scaled scores, all of the item separation indices, both sample-based and model-based, were greater than 220. For the mathematics scaled scores, all of the item separation indices, both sample-based and model-based, were greater than 180. Taken together, these findings indicate it is reasonable to expect highly consistent estimates of item difficulty levels across samples. The person-based reliability coefficients are outlined below by domain of development.

Social–Emotional Scale

Based on the Rasch reliability indices, the scaled scores appear to yield highly reliable information from this sample, as evidenced by the sample-based person separation index (7.43) and model-based person separation index (8.38). Similarly, the sample-based person reliability index (.98) and the model-based person reliability (.99) were high. The Cronbach's alpha value was .99, indicating high internal consistency reliability.

Physical Scale

Based on the Rasch reliability indices, the scaled scores appear to yield highly reliable information from this sample, as evidenced by the sample-based person separation index (8.06) and model-based person separation index (7.09). Similarly, the sample-based person reliability index (.97) and the model-based person reliability (.98) were high. The Cronbach's alpha value was .98, indicating high internal consistency reliability.

Language Scale

Based on the Rasch reliability indices, the scaled scores appear to yield highly reliable information from this sample, as evidenced by the sample-based person separation index (8.06) and model-based person separation index (9.10). Similarly, the sample-based person reliability index (.98) and the model-based person reliability (.99) were high. The Cronbach's alpha value was .99, indicating high internal consistency reliability.

Cognitive Scale

Based on the Rasch reliability indices, the scaled scores appear to yield highly reliable information from this sample, as evidenced by the sample-based person separation index (8.87) and model-based person separation index (9.94). Similarly, the sample-based person reliability index (.99) and the model-based person reliability (.99) were high. The Cronbach's alpha value was .99, indicating high internal consistency reliability.

Literacy Scale

Based on the Rasch reliability indices, the scaled scores appear to yield highly reliable information from this sample, as evidenced by the sample-based person separation index (4.69) and model-based person separation index (5.36). Similarly, the sample-based person reliability index (.96) and the model-based person reliability (.97) were high. The Cronbach's alpha value was .98, indicating high internal consistency reliability.

Mathematics Scale

Based on the Rasch reliability indices, the scaled scores appear to yield highly reliable information from this sample, as evidenced by the sample-based person separation index (5.93) and model-based person separation index (6.54). Similarly, the sample-based person reliability index (.97) and the model-based person reliability (.98) were high. The Cronbach's alpha value was .97, indicating high 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 young children across all six domains of development. The Rasch reliability indices were greater than .95 across all scales, and the person separation indices were all greater than 4.00.

Summary

The primary goal of this study was to extend the reliability and validity evidence for *GOLD*® using a nationally representative sample of children served by teachers who assess children in fall, winter, and spring of the academic year. Overall, this study reaffirmed that the use of *GOLD*® with young children continues to yield highly reliable scaled scores as indicated by both the classical and Rasch reliability statistics. The results also demonstrate strong statistical evidence that the developmental progressions within each scale generally work very well together to measure a single underlying domain of development. The results further demonstrate evidence that the ratings can be successfully organized within each developmental domain generally as intended by the team that originally developed the assessment. There is also statistical evidence that teachers are able to use the rating scale effectively to place children along a progression of development and learning. When the items within each domain are arranged from the easier to the most difficult objectives for children to master, the hierarchy that is created generally 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 second- and third-grade children is needed for future studies, especially given the relatively small numbers of children in these grade levels in the study, and the relatively smaller numbers of children placed at the upper ends of the many of the progressions. Future research could focus on measures of the degree of association between *GOLD*® scaled scores and external measures of child developmental progress. It would also be helpful to conduct a study that addresses interrater reliability. Follow-up studies should also 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 assessment, 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 differences between subgroups of children. Lastly, future research is needed to evaluate whether teachers are collecting sufficient quantities of high quality, valid evidence of child growth, development, and learning to support placements on the *GOLD*® developmental progressions.

Teachers are being asked to implement a variety of assessments and assessment-related tasks. They must do so while facing challenges related to limited instructional resources, many demands on their time, and increasingly diverse classrooms of children with individualized needs. All of this comes at a time when teachers are often unprepared for the linguistic and cultural diversity that is the reality in most classrooms. To plan instruction accordingly, they need assessments to help them document and understand the developmental status, strengths, needs, and growth patterns of their children. This study has robustly demonstrated that *GOLD*® satisfies the requirement for high-quality assessment in early childhood education programs.

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. (2017). *Technical manual for the* Teaching Strategies GOLD® *assessment system: Birth through third grade edition*. Technical Report. Charlotte, N.C.: *Center for Educational Measurement and Evaluation*, University of North Carolina Charlotte.

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., Kim, D-H., & Burts, D. (2013). *Technical manual for the* Teaching Strategies GOLD® *assessment system (second edition)*. Technical Report. Charlotte, N.C.: *Center forEducational Measurement and Evaluation*, University of North Carolina Charlotte.

Lambert, R., Kim, D-H., & Burts, D. (2014). *Technical manual for the* Teaching Strategies GOLD® *assessment system (third edition)*. Technical Report. Charlotte, N.C.: *Center for Educational Measurement and Evaluation*, University of North Carolina Charlotte.

Lambert, R., Kim, D-H., & Burts, D. (2013). *Evidence for the Association between Scores from the* Teaching Strategies GOLD® *Assessment System and Information from Direct Assessments of Child Progress*. 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. (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
Norm sample by age/grade

Age/grade	n	%
Birth to 1 year	5,000	15.60%
1 to 2 years	5,000	15.60%
2 to 3 years	5,000	15.60%
Preschool 3	5,000	15.60%
Prekindergarten 4	5,000	15.60%
Kindergarten	5,000	15.60%
First grade	1,626	5.10%
Second grade	377	1.20%
Third grade	60	0.20%
Total	**31,626**	**100.00%**

Table 2
Demographic characteristics

Characteristic	Category	%
Gender	Female	51.48%
	Male	48.52%
Race/Ethnicity	White	49.69%
	African American	13.89%
	Nt. Am. & Pac. Is.	1.22%
	Asian	4.13%
	Multiple Races	4.64%
	Hispanic	26.43%
Primary Language	English	78.57%
	Spanish	14.47%
	Other	6.96%
Disability Status	Has IFSP or IEP	7.24%
Poverty Status	Qualifies for FRL	33.19%

Table 3
Reliability coefficients for all scales

Domain	Index	Coefficient
Social–Emotional	Cronbach's alpha	0.99
	Sample-based Person Separation Index	7.43
	Sample-based Person Reliability	0.98
	Model-based Person Separation Index	8.38
	Model-based Person Reliability	0.99
Physical	Cronbach's alpha	0.98
	Sample-based Person Separation Index	6.18
	Sample-based Person Reliability	0.97
	Model-based Person Separation Index	7.09
	Model-based Person Reliability	0.98
Language	Cronbach's alpha	0.99
	Sample-based Person Separation Index	8.06
	Sample-based Person Reliability	0.98
	Model-based Person Separation Index	9.10
	Model-based Person Reliability	0.99
Cognitive	Cronbach's alpha	0.99
	Sample-based Person Separation Index	8.87
	Sample-based Person Reliability	0.99
	Model-based Person Separation Index	9.94
	Model-based Person Reliability	0.99
Literacy	Cronbach's alpha	0.98
	Sample-based Person Separation Index	4.69
	Sample-based Person Reliability	0.96
	Model-based Person Separation Index	5.36
	Model-based Person Reliability	0.97
Mathematics	Cronbach's alpha	0.97
	Sample-based Person Separation Index	5.93
	Sample-based Person Reliability	0.97
	Model-based Person Separation Index	6.54
	Model-based Person Reliability	0.98

Table 4

Fall item-level statistics and difficulty estimates for the social–emotional scale

					Item-Measure r	
Progression	Item Difficulty	SE	Infit Mnsq	Outfit Mnsq	Item Included	Item Excluded
3.A	1.29	0.01	0.86	0.83	0.95	0.94
3.B	1.21	0.01	0.96	0.95	0.95	0.95
2.D	0.78	0.01	0.96	1.03	0.94	0.94
2.C	0.59	0.01	0.85	0.84	0.94	0.93
2.B	0.49	0.01	0.88	0.88	0.94	0.94
1.A	-0.12	0.01	1.12	1.11	0.92	0.93
1.B	-0.60	0.01	0.89	0.88	0.95	0.94
1.C	-0.85	0.01	0.94	0.93	0.95	0.95
2.A	-2.80	0.01	1.18	1.20	0.92	0.93

Table 5

Fall item-level statistics and difficulty estimates for the physical scale

					Item-Measure r	
Progression	Item Difficulty	SE	Infit Mnsq	Outfit Mnsq	Item Included	Item Excluded
7.B	1.27	0.01	1.22	2.22	0.94	0.95
5	0.41	0.01	0.91	0.90	0.95	0.95
6	0.38	0.01	0.84	0.81	0.95	0.94
4	-1.02	0.01	0.84	0.82	0.96	0.95
7.A	-1.03	0.01	0.88	0.84	0.95	0.95

Table 6

Fall item-level statistics and difficulty estimates for the language scale

					Item-Measure r	
Progression	Item Difficulty	SE	Infit Mnsq	Outfit Mnsq	Item Included	Item Excluded
9.D	1.06	0.01	0.92	0.92	0.94	0.94
10.B	0.71	0.01	1.02	1.03	0.95	0.95
9.C	0.52	0.01	0.87	0.89	0.96	0.96
10.A	0.32	0.01	0.89	0.91	0.96	0.96
9.A	0.25	0.01	0.78	0.79	0.96	0.95
8.A	0.01	0.01	0.90	0.90	0.96	0.95
9.B	-0.07	0.01	0.95	0.97	0.96	0.96
8.B	-2.80	0.01	1.15	1.17	0.95	0.96

Table 7

Fall item-level statistics and difficulty estimates for the cognitive scale

					Item-Measure r	
Progression	Item Difficulty	SE	Infit Mnsq	Outfit Mnsq	Item Included	Item Excluded
12.B	0.97	0.01	0.79	0.79	0.96	0.96
11.E	0.94	0.01	0.88	0.90	0.95	0.95
11.D	0.55	0.01	0.89	0.91	0.95	0.95
12.A	0.31	0.01	0.89	0.90	0.95	0.95
14.B	0.23	0.01	1.00	1.04	0.95	0.95
14.A	-0.06	0.01	0.90	0.97	0.95	0.95
13	-0.06	0.01	1.00	1.12	0.95	0.95
11.A	-0.33	0.01	1.07	1.06	0.93	0.93
11.C	-0.96	0.01	0.96	0.96	0.96	0.96
11.B	-1.60	0.01	0.95	0.96	0.95	0.95

Table 8

Fall item-level statistics and difficulty estimates for the literacy scale

					Item-Measure r	
Progression	Item Difficulty	SE	Infit Mnsq	Outfit Mnsq	Item Included	Item Excluded
18.D	2.99	0.01	1.30	4.78	0.55	0.56
19.C	2.97	0.01	1.20	1.42	0.56	0.56
18.E	2.90	0.01	0.98	0.95	0.55	0.55
15.D	1.80	0.01	1.02	1.39	0.64	0.65
18.C	1.07	0.01	0.85	0.82	0.82	0.81
15.C	0.94	0.01	0.91	2.03	0.76	0.76
18.A	0.78	0.01	0.84	0.82	0.85	0.84
19.B	0.10	0.01	1.31	1.18	0.85	0.85
17.A	-0.32	0.01	0.86	0.95	0.90	0.90
16.B	-1.01	0.01	0.89	1.02	0.74	0.74
17.B	-1.19	0.01	0.70	0.77	0.83	0.82
15.A	-1.68	0.01	1.02	1.05	0.85	0.86
16.A	-2.10	0.01	1.13	1.22	0.81	0.81
18.B	-2.18	0.01	0.70	0.68	0.84	0.83
15.B	-2.20	0.01	0.80	0.79	0.84	0.83
19.A	-2.88	0.00	1.47	1.35	0.85	0.86

Table 9

Fall item-level statistics and difficulty estimates for the mathematics scale

					Item-Measure r	
Progression	Item Difficulty	SE	Infit Mnsq	Outfit Mnsq	Item Included	Item Excluded
20.E	2.37	0.01	0.93	0.98	0.60	0.60
20.D	2.27	0.01	1.12	2.19	0.56	0.57
20.F	2.01	0.01	1.12	1.05	0.59	0.59
22.C	1.04	0.01	1.11	1.06	0.78	0.78
22.B	0.79	0.01	1.37	1.83	0.82	0.84
22.A	-0.14	0.01	1.20	1.06	0.89	0.90
23	-0.71	0.01	0.90	0.94	0.92	0.92
21.B	-0.75	0.01	0.95	0.92	0.94	0.93
20.A	-0.87	0.01	0.78	0.79	0.94	0.93
20.B	-0.92	0.01	0.69	0.71	0.92	0.91
20.C	-0.98	0.01	0.84	0.77	0.88	0.87
21.A	-4.11	0.01	1.06	1.12	0.94	0.94

Table 10

Social-emotional scaled scores by age/grade

				Fall to Spring	
Age / grade		Fall	Winter	Spring	Gain
Birth to 1 year	Mean	165.98	209.26	247.70	82.43
	SEM	21	20	17	
	SD	59.34	55.78	54.44	54.23
	25th	126	168	219	49
	50th	168	208	250	82
	75th	208	240	282	114
1 to 2 years	Mean	260.03	292.30	314.07	54.04
	SEM	17	16	15	
	SD	59.39	52.40	50.89	54.20
	25th	230	258	282	22
	50th	267	297	318	51
	75th	297	325	343	82
2 to 3 years	Mean	329.74	354.90	373.88	44.29
	SEM	14	13	13	
	SD	60.02	58.34	59.08	51.01
	25th	305	325	343	16
	50th	331	354	376	40
	75th	365	387	403	70
Preschool 3	Mean	379.88	416.85	442.81	63.30
	SEM	13	14	13	
	SD	61.98	59.17	65.57	57.43
	25th	349	387	409	31
	50th	381	420	442	58
	75th	415	447	478	89
Prekindergarten 4	Mean	426.44	466.51	497.35	70.74
	SEM	13	13	13	
	SD	55.57	54.07	62.53	56.02
	25th	398	442	467	39
	50th	431	467	499	66
	75th	457	499	527	96
Kindergarten	Mean	473.23	522.66	558.00	80.87
	SEM	13	14	14	
	SD	54.59	55.79	60.79	49.05
	25th	447	494	527	52
	50th	483	527	569	82
	75th	505	556	593	106
1st grade	Mean	560.50	601.18	639.25	72.16
	SEM	14	14	13	
	SD	72.19	73.35	68.07	51.79
	25th	533	569	611	46
	50th	587	617	656	69
	75th	605	650	674	94

Table 11

Physical scaled scores by age/grade

				Fall to Spring	
Age / grade		Fall	Winter	Spring	Gain
Birth to 1 year	Mean	184.19	254.43	311.33	128.01
	SEM	28	28	21	
	SD	86.95	77.79	73.14	74.99
	25th	141	215	270	81
	50th	189	244	320	128
	75th	244	306	347	172
1 to 2 years	Mean	337.19	377.51	405.44	68.20
	SEM	20	22	23	
	SD	80.31	69.99	70.88	73.73
	25th	290	334	361	28
	50th	334	376	411	67
	75th	393	411	444	106
2 to 3 years	Mean	430.96	462.68	484.84	53.64
	SEM	22	20	19	
	SD	80.00	76.53	76.86	70.67
	25th	393	411	444	14
	50th	428	470	494	50
	75th	482	506	530	89
Preschool 3	Mean	496.07	537.37	567.59	71.79
	SEM	19	20	20	
	SD	76.33	69.46	74.54	71.10
	25th	458	506	530	30
	50th	506	543	568	63
	75th	543	568	604	105
Prekindergarten 4	Mean	547.91	593.33	628.50	80.73
	SEM	20	19	18	
	SD	66.60	60.64	65.00	65.89
	25th	518	556	593	43
	50th	556	593	635	76
	75th	581	625	668	112
Kindergarten	Mean	602.15	655.72	693.14	92.51
	SEM	18	19	18	
	SD	62.28	57.23	61.70	56.14
	25th	568	635	668	61
	50th	615	668	710	93
	75th	635	680	726	122
1st grade	Mean	683.07	727.36	749.89	61.09
	SEM	19	16	15	
	SD	72.34	53.40	53.01	50.87
	25th	668	710	734	30
	50th	710	742	757	54
	75th	726	757	771	78

Table 12
Language scaled scores by age/grade

				Fall to Spring	
Age / grade		Fall	Winter	Spring	Gain
Birth to 1 year	Mean	128.49	172.95	214.09	85.79
	SEM	21	20	21	
	SD	63.35	62.01	62.98	59.44
	25th	100	131	183	53
	50th	131	170	211	83
	75th	170	211	248	119
1 to 2 years	Mean	236.29	274.99	305.59	69.38
	SEM	19	17	16	
	SD	70.53	68.01	67.07	61.65
	25th	197	236	268	34
	50th	248	277	310	66
	75th	286	318	348	105
2 to 3 years	Mean	341.46	372.97	397.38	55.64
	SEM	16	16	16	
	SD	78.61	76.10	77.14	59.66
	25th	302	333	356	23
	50th	348	372	395	52
	75th	387	417	445	86
Preschool 3	Mean	402.14	442.82	474.07	72.40
	SEM	15	15	16	
	SD	82.86	80.79	87.40	67.11
	25th	364	403	431	35
	50th	410	445	473	65
	75th	452	489	524	104
Prekindergarten 4	Mean	461.29	507.14	544.96	84.08
	SEM	15	17	17	
	SD	72.25	71.42	78.74	63.56
	25th	424	473	506	45
	50th	473	515	551	79
	75th	506	551	593	117
Kindergarten	Mean	516.61	570.84	610.39	86.42
	SEM	17	17	17	
	SD	74.44	71.43	75.56	59.31
	25th	473	542	576	50
	50th	524	585	627	85
	75th	560	619	659	119
1st grade	Mean	619.42	665.67	699.13	79.69
	SEM	17	15	13	
	SD	71.99	87.37	76.15	44.06
	25th	593	636	673	54
	50th	644	679	718	80
	75th	666	707	746	105

Table 13
Cognitive scaled scores by age/grade

				Fall to Spring	
Age / grade		Fall	Winter	Spring	Gain
Birth to 1 year	Mean	126.48	171.77	211.76	85.96
	SEM	20	19	16	
	SD	62.27	57.26	55.97	54.39
	25th	75	132	184	52
	50th	132	172	215	85
	75th	172	206	240	117
1 to 2 years	Mean	229.40	264.43	290.24	60.74
	SEM	15	14	14	
	SD	60.83	55.22	55.55	53.55
	25th	196	232	262	29
	50th	232	269	298	58
	75th	269	298	326	90
2 to 3 years	Mean	316.50	346.01	368.33	51.63
	SEM	14	13	13	
	SD	65.95	63.27	65.38	55.41
	25th	284	312	332	20
	50th	319	351	369	49
	75th	357	381	401	79
Preschool 3	Mean	374.62	415.58	444.42	70.38
	SEM	13	14	13	
	SD	68.85	66.08	71.39	60.36
	25th	338	381	408	34
	50th	381	415	448	66
	75th	415	454	481	98
Prekindergarten 4	Mean	427.94	472.29	506.46	78.79
	SEM	14	12	13	
	SD	63.36	58.58	65.12	59.66
	25th	395	442	470	44
	50th	442	475	509	74
	75th	465	509	545	108
Kindergarten	Mean	480.12	536.18	575.75	89.87
	SEM	12	13	12	
	SD	65.96	64.60	70.15	56.72
	25th	448	503	533	59
	50th	492	545	592	92
	75th	521	578	622	119
1st grade	Mean	598.08	637.09	663.00	65.08
	SEM	11	11	11	
	SD	77.92	67.21	79.32	54.55
	25th	573	622	646	37
	50th	626	650	679	63
	75th	642	671	705	92

Table 14
Literacy scaled sores by age/grade

				Fall to Spring	
Age / grade		Fall	Winter	Spring	Gain
Birth to 1 year	Mean	86.97	146.50	211.85	125.39
	SEM	82	68	45	
	SD	95.61	101.55	101.68	102.48
	25th	0	110	110	64
	50th	110	110	207	110
	75th	110	207	290	207
1 to 2 years	Mean	249.56	298.94	330.13	80.05
	SEM	34	24	21	
	SD	101.70	83.49	77.15	88.88
	25th	207	271	305	28
	50th	271	305	341	68
	75th	318	351	377	123
2 to 3 years	Mean	365.48	392.36	412.42	46.60
	SEM	18	16	15	
	SD	78.87	68.58	67.13	66.49
	25th	341	360	385	15
	50th	377	398	425	41
	75th	410	435	450	72
Preschool 3	Mean	428.30	461.74	481.08	53.18
	SEM	13	11	9	
	SD	66.18	54.71	55.11	52.94
	25th	404	443	460	27
	50th	443	470	487	47
	75th	467	491	509	71
Prekindergarten 4	Mean	474.47	506.52	527.01	52.52
	SEM	10	8	8	
	SD	51.10	39.56	43.10	44.00
	25th	457	491	508	31
	50th	485	509	531	48
	75th	502	529	550	67
Kindergarten	Mean	525.74	570.16	602.01	71.22
	SEM	8	8	9	
	SD	45.20	47.37	54.02	38.99
	25th	506	547	573	48
	50th	529	573	606	70
	75th	550	598	634	92
1st grade	Mean	615.95	660.78	692.92	73.85
	SEM	10	11	12	
	SD	51.28	57.81	53.93	33.25
	25th	591	634	673	54
	50th	624	667	700	71
	75th	646	689	721	91

Table 15
Mathematics scaled scores by age/grade

				Fall to Spring	
Age / grade		Fall	Winter	Spring	Gain
Birth to 1 year	Mean	10.87	24.41	58.31	47.95
	SEM	80	63	46	
	SD	38.65	51.61	69.44	61.67
	25th	0	0	0	0
	50th	0	0	56	17
	75th	0	56	93	93
1 to 2 years	Mean	93.29	137.30	174.45	80.98
	SEM	35	27	24	
	SD	77.92	77.53	74.85	71.21
	25th	0	93	136	37
	50th	93	153	181	80
	75th	153	194	229	125
2 to 3 years	Mean	217.47	248.81	272.17	54.85
	SEM	22	20	18	
	SD	75.01	68.31	67.61	59.07
	25th	181	218	239	22
	50th	229	257	280	49
	75th	265	292	315	84
Preschool 3	Mean	290.08	327.62	352.05	62.56
	SEM	16	15	14	
	SD	66.47	58.31	58.42	51.34
	25th	265	304	325	33
	50th	298	331	354	56
	75th	331	359	384	86
Prekindergarten 4	Mean	341.41	379.14	405.25	63.96
	SEM	14	13	13	
	SD	53.48	47.28	51.44	46.33
	25th	320	354	380	37
	50th	350	384	408	59
	75th	376	408	433	83
Kindergarten	Mean	385.08	440.51	481.27	95.32
	SEM	13	13	13	
	SD	57.55	53.04	56.26	51.79
	25th	354	413	454	65
	50th	392	446	490	91
	75th	421	474	514	123
1st grade	Mean	471.07	532.19	583.44	108.99
	SEM	13	13	14	
	SD	65.82	66.48	63.46	55.19
	25th	432	486	557	75
	50th	486	548	599	103
	75th	514	575	622	132

Table 16
Widely held expectations for social-emotional by age/grade

		Fall %	Winter %	Spring %
Birth to 1 year	Below Expectations	9.3	2.0	0.8
	Meets Expectations	86.1	84.2	61.7
	Exceeds Expectations	4.6	13.8	37.5
1 to 2 years	Below Expectations	30.0	11.6	5.7
	Meets Expectations	67.0	80.3	78.2
	Exceeds Expectations	3.0	8.0	16.1
2 to 3 years	Below Expectations	33.9	18.3	11.9
	Meets Expectations	59.5	68.5	65.3
	Exceeds Expectations	6.6	13.2	22.9
Preschool 3	Below Expectations	39.8	17.8	9.4
	Meets Expectations	53.9	65.7	59.4
	Exceeds Expectations	6.2	16.4	31.1
Prekindergarten 4	Below Expectations	55.4	24.7	13.3
	Meets Expectations	42.6	65.6	58.6
	Exceeds Expectations	2.0	9.7	28.0
Kindergarten	Below Expectations	59.3	22.2	11.2
	Meets Expectations	40.1	74.3	74.0
	Exceeds Expectations	0.7	3.5	14.8
1st grade	Below Expectations	63.5	35.6	17.7
	Meets Expectations	35.3	56.9	61.4
	Exceeds Expectations	1.2	7.5	21.0

Table 17
Widely held expectations for physical by age/grade

		Fall %	Winter %	Spring %
Birth to 1 year	Below Expectations	21.2	4.3	0.8
	Meets Expectations	74.9	81.7	57.3
	Exceeds Expectations	3.9	14.1	41.9
1 to 2 years	Below Expectations	32.4	13.4	7.2
	Meets Expectations	60.0	71.1	64.0
	Exceeds Expectations	7.6	15.5	28.8
2 to 3 years	Below Expectations	28.0	15.6	9.7
	Meets Expectations	62.6	67.2	63.2
	Exceeds Expectations	9.4	17.2	27.1
Preschool 3	Below Expectations	31.0	13.6	7.4
	Meets Expectations	63.3	71.5	62.5
	Exceeds Expectations	5.7	14.9	30.1
Prekindergarten 4	Below Expectations	43.6	18.8	9.1
	Meets Expectations	55.0	75.0	71.7
	Exceeds Expectations	1.4	6.2	19.2
Kindergarten	Below Expectations	57.8	18.5	8.3
	Meets Expectations	40.9	76.6	70.7
	Exceeds Expectations	1.2	4.9	21.0
1st grade	Below Expectations	46.6	19.6	8.4
	Meets Expectations	52.5	72.4	69.0
	Exceeds Expectations	0.9	8.0	22.6

Table 18
Widely held expectations for language by age/grade

		Fall %	Winter %	Spring %
Birth to 1 year	Below Expectations	21.8	7.1	3.1
	Meets Expectations	76.6	88.3	78.1
	Exceeds Expectations	1.7	4.6	18.7
1 to 2 years	Below Expectations	50.0	27.0	14.4
	Meets Expectations	48.9	68.8	75.1
	Exceeds Expectations	1.2	4.2	10.4
2 to 3 years	Below Expectations	42.6	26.7	17.2
	Meets Expectations	51.2	60.9	61.4
	Exceeds Expectations	6.1	12.4	21.4
Preschool 3	Below Expectations	46.3	26.1	16.5
	Meets Expectations	49.9	63.4	61.4
	Exceeds Expectations	3.8	10.5	22.1
Prekindergarten 4	Below Expectations	49.8	24.0	13.8
	Meets Expectations	48.6	70.4	67.4
	Exceeds Expectations	1.6	5.6	18.7
Kindergarten	Below Expectations	67.3	32.6	17.3
	Meets Expectations	31.8	62.7	64.2
	Exceeds Expectations	0.9	4.7	18.5
1st grade	Below Expectations	64.5	36.1	17.1
	Meets Expectations	35.2	60.4	68.9
	Exceeds Expectations	0.3	3.5	14.0

Table 19
Widely held expectations for cognitive by age/grade

		Fall %	Winter %	Spring %
Birth to 1 year	Below Expectations	13.2	2.9	0.9
	Meets Expectations	83.4	86.1	63.7
	Exceeds Expectations	3.4	11.1	35.4
1 to 2 years	Below Expectations	17.9	5.3	2.5
	Meets Expectations	78.7	85.7	76.4
	Exceeds Expectations	3.4	9.0	21.1
2 to 3 years	Below Expectations	31.8	16.5	9.9
	Meets Expectations	61.7	70.1	67.0
	Exceeds Expectations	6.5	13.4	23.1
Preschool 3	Below Expectations	45.9	22.6	13.6
	Meets Expectations	50.2	64.9	61.7
	Exceeds Expectations	3.9	12.5	24.7
Prekindergarten 4	Below Expectations	54.3	25.0	12.7
	Meets Expectations	44.8	71.0	73.1
	Exceeds Expectations	0.9	3.9	14.3
Kindergarten	Below Expectations	71.3	32.1	17.0
	Meets Expectations	28.1	66.6	74.5
	Exceeds Expectations	0.5	1.4	8.5
1st grade	Below Expectations	70.5	36.8	20.3
	Meets Expectations	27.7	59.9	64.2
	Exceeds Expectations	1.8	3.3	15.5

Table 20
Widely held expectations for literacy by age/grade

		Fall %	Winter %	Spring %
Birth to 1 year	Below Expectations	0.0	0.0	0.0
	Meets Expectations	92.8	80.0	52.2
	Exceeds Expectations	7.2	20.0	47.8
1 to 2 years	Below Expectations	18.8	6.4	3.3
	Meets Expectations	76.8	83.2	74.5
	Exceeds Expectations	4.4	10.4	22.2
2 to 3 years	Below Expectations	40.3	25.1	16.1
	Meets Expectations	53.4	63.0	63.7
	Exceeds Expectations	6.3	11.8	20.2
Preschool 3	Below Expectations	53.6	25.9	14.8
	Meets Expectations	42.4	63.4	61.4
	Exceeds Expectations	3.9	10.7	23.8
Prekindergarten 4	Below Expectations	52.5	21.1	10.1
	Meets Expectations	46.5	75.0	75.0
	Exceeds Expectations	1.0	3.9	15.0
Kindergarten	Below Expectations	52.7	14.9	7.2
	Meets Expectations	46.8	80.0	69.9
	Exceeds Expectations	0.5	5.1	22.9
1st grade	Below Expectations	73.0	30.2	13.7
	Meets Expectations	25.4	61.3	61.8
	Exceeds Expectations	1.6	8.5	24.5

Table 21
Widely held expectations for mathematics by age/grade

		Fall %	Winter %	Spring %
Birth to 1 year	Below Expectations	0.0	0.0	0.0
	Meets Expectations	97.3	93.1	76.6
	Exceeds Expectations	2.7	6.9	23.4
1 to 2 years	Below Expectations	43.1	21.8	10.5
	Meets Expectations	54.8	72.6	74.2
	Exceeds Expectations	2.1	5.6	15.2
2 to 3 years	Below Expectations	39.6	23.7	14.1
	Meets Expectations	55.5	66.3	67.9
	Exceeds Expectations	4.9	10.0	18.0
Preschool 3	Below Expectations	45.6	20.7	10.8
	Meets Expectations	49.6	65.0	60.7
	Exceeds Expectations	4.7	14.2	28.6
Prekindergarten 4	Below Expectations	69.6	36.0	18.5
	Meets Expectations	29.8	59.6	65.7
	Exceeds Expectations	0.6	4.4	15.7
Kindergarten	Below Expectations	78.7	33.8	13.2
	Meets Expectations	20.8	63.7	73.0
	Exceeds Expectations	0.5	2.5	13.8
1st grade	Below Expectations	80.1	35.1	14.6
	Meets Expectations	19.9	60.9	72.5
	Exceeds Expectations	0.0	4.0	12.9

Glossary

Colored Band – Under each developmental progression, colored bands provide the range of widely held expectations for a particular group (age, class, or grade) of children's knowledge, skills, and abilities.

Developmental Progression – Each developmental progression provides a continuum of behavioral expectations under a given dimension for birth through third grade. Teachers rate children's abilities according to the scale provided and the evidence gathered for each child.

Dimension – Dimensions are the narrowest levels of *GOLD*®'s hierarchy. Objectives are made up of one or more dimensions that further define the associated observable behaviors.

Domain – Domains are broadest levels of *GOLD*®'s hierarchy. They are the comprehensive areas of development and learning that teachers assess using *GOLD*®. The ten domains include four developmental domains: social-emotional, physical, language, and cognitive; five content domains: literacy, mathematics, science and technology, social studies, and the arts; and one additional domain for English and dual-language learners: English language acquisition.

Objective – Each domain is made up of a set of objectives designed to guide teachers through the assessment process. Objectives enable teachers to link observable behavior to essential developmental and early learning requirements. *GOLD*® has thirty-eight objectives in total, which Teaching Strategies calls the Objectives for Development and Learning (ODL).

Raw Score – A raw score represents the placement of a child's knowledge, skills, and abilities, or current status level, on a given developmental progression. Summing the raw scores within a domain will provide a child's total raw score for that domain.

Scaled Score – Scaled scores convert total raw scores to a common scale. These scores can be used to track children's growth and development over time. The *GOLD*® uniform scale ranges from 0 to 1000 for six areas of development and learning: social-emotional, physical, language, cognitive, literacy, and mathematics. Unlike with raw scores, with scaled scores one point of growth represents the "same amount" of development within a given domain.

Widely Held Expectations (WHEs) – WHEs are the range of knowledge, skills, and abilities that children of a particular age, class, or grade typically demonstrate over a given year of life or throughout a program/school year.

TeachingStrategies.com