CEMETR-2014-06
NOVEMBER 2014

CEME Technical Report

The Center for Educational Measurement and Evaluation

Technical Manual (3rd Edition)
for the *Teaching Strategies GOLD*
Assessment System

Richard Lambert
Do-Hong Kim
Diane Burts

RICHARD LAMBERT
CHUANG WANG
MARK D'AMICO
EDITORS

A PUBLICATION OF
THE CENTER FOR
EDUCATIONAL
MEASUREMENT
AND EVALUATION

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

Technical Manual for the Teaching Strategies *GOLD™* Assessment System

(3rd edition)

Richard G. Lambert, Ph.D.

Do-Hong Kim, Ph.D.

Center for Educational Measurement and Evaluation

UNC Charlotte

Diane C. Burts, Ed.D.

Louisiana State University

November, 2014

Teaching Strategies GOLD™ Assessment System (GOLD) is a formative assessment system that has been designed and validated for use with young children ages birth to kindergarten. The 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. The GOLD system 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 young children.

Background Information on the Development of GOLD

GOLD (Heroman et al., 2010) measures the progress of children ages birth through kindergarten 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 behaviors 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 23 GOLD objectives included in the current studies are operationalized into 51 rating scale items: social–emotional (9 items), physical (5 items), language (8 items), cognitive (10 items), literacy (12 items) and mathematics (7 items). Teachers rate children's skills, knowledge, and behaviors along a 10- point progression of development and learning from "Not Yet" (Level 0) to Level 9 (exceeds kindergarten expectations) using collected documentation evidence. Levels 2, 4, 6, and 8 are "Indicators" and include varied examples from everyday situations that give teachers guidance of what the evidence may look like with majority and with subgroups of children. Levels 1, 3, 5, 7, and 9 are "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 (αs = .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 to more than 45,000, with over a 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.

National Norm Sample

From the total population of children assessed using GOLD, a sample was selected to be nationally representative with respect to ethnicity. The first step in creating the national norm sample was to screen the data for valid birth dates and assessment dates. Admissible data was defined as containing birthdates that indicated valid child ages in months at the beginning of the academic year for the type of classroom in which the child was placed. The six classroom types or age / grade bands are: 1.) infants, 2.) one-year-olds, 3.) two-year-olds, 4.) three-year-olds, 5.) four-year-olds, and 6.) kindergarten. In addition, children had to have valid complete assessment data for fall, winter, and spring checkpoints, with the exception of kindergarten where many schools and programs use GOLD for kindergarten entry assessment only. The timing of the fall, winter, and spring assessments had to be within specific time periods to eliminate data from programs that use unconventional checkpoints or non-traditional schedules. After reducing the population to cases that met these criteria, stratified random sampling, stratifying on ethnicity and age, was used to select 5,000 children from each of the six age / grade bands. The primary sampling unit was the child, not the classroom, to minimize clustering and rater effects.

The 2013 Census Bureau national estimates for the proportion of children ages birth to 6 years of age in each ethnicity / race group were used to set the proportional allocation targets. Teachers are required to enter into the GOLD online system information regarding each child's race and ethnicity. The questions about each child are the same as 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, not Hispanic; 4.) Asian, not Hispanic; 5.) Hawaiian / Pacific Islander, not Hispanic; 6.) multiracial, not Hispanic; and 7.) Hispanic.

As shown in Table 1, a total of 30,000 children were retained in the norm sample. These children received educational services in centers or schools that were located in all regions of the United States. These programs and centers included Head Start, private childcare, and school-based sites. All fifty states, Puerto Rico, and the District of Columbia were represented in each of the six age / grade bands. The percentage of the norm sample from each race and ethnicity group very nearly replicated the national Census Bureau 2013 estimates. The only exceptions were for White children who were slightly over represented and Asian children who were slightly under represented. This result was related to the fact that Asian children were under represented in the GOLD population of infants assessed.

Across all the children in the norm sample, the fall assessment took place, on average, 2.67 months after the beginning of academic year (SD=.67). The winter assessment took place, on average, 6.01 months after the beginning of academic year (SD=.70). The spring assessment took place, on average, 9.64 months after the beginning of academic year (SD=1.13). As shown in Table 2, the norm sample was very evenly balanced by gender (boys=51.2%, girls=48.8%). Children with an IEP or IFSP comprised 9.5% of the norm sample. A total of 27.4% of the norm sample qualified for free or reduced price lunch. The primary language spoken in the home was distributed as follows: English (79.4%), Spanish (14.9%), and other languages (5.8%).

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 mostly share the same rating scale structure (i.e., use of the same number of rating scale categories and labels across items), however a small subset of the items uses a slightly different rating scale. In cases where each item has its own rating scale structure, the PCM is the appropriate model to apply.

Dimensionality

Rasch modeling assumes what is called unidimensionality, meaning that the items in question measure one and 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). For PCAR, a variance of greater than 50% explained by measures is considered good, supporting 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).

Cognitive Scale (10 items)

The PCAR showed that for the Cognitive scale, the Rasch dimension explained 90.1% of the variance in the data, with its eigenvalue of 91.0. The first contrast (the largest secondary dimension) had an eigenvalue of 2.2 and accounted only for 2.2% of the unexplained variance. The fit statistics

for all of the Cognitive items were within acceptable limits: the infit MNSQ ranged from 0.83 to 1.22; the outfit MNSQ ranged from 0.85 to 1.20. The item total score correlations ranged from .93 to .95.

Language Scale (8 items)

The PCAR showed that for the Language scale, the Rasch dimension explained 91.3% of the variance in the data, with its eigenvalue of 83.9. The first contrast (the largest secondary dimension) had an eigenvalue of 1.9 and accounted only for 2.1% of the unexplained variance. The fit statistics for all of the Language items were well within acceptable limits: the infit MNSQ ranged from 0.73 to 1.09; the outfit MNSQ ranged from 0.83 to 1.18. The item total score correlations ranged from .91 to .95.

Literacy Scale (12 items)

The PCAR showed that the Rasch dimension explained 85.9% of the variance in the data, with its eigenvalue of 73.0. The first contrast (the largest secondary dimension) had an eigenvalue of 2.3 and accounted for 2.7% of the unexplained variance. The fit statistics for the Literacy items were mostly within acceptable limits: the infit MNSQ ranged from 0.69 to 1.94; the outfit MNSQ ranged from 0.60 to 1.53. Item 16.A (identifies and names letters) yielded MNSQ statistics that were beyond the acceptable range (1.94 and 1.53). Items with mean square values of between 1.5-2.0 can be considered unproductive for the construction of measurement scales, but not degrading to the quality of the information provided by the scale (Linacre, 2002).

This item did, however, yield an item total score correlation of .88, illustrating that it does provide information that is related to the rest of the information provided by this set of items. The item total score correlations ranged from .88 to .93.

Mathematics Scale (7 items)

The PCAR showed that the Rasch dimension explained 88.0% of the variance in the data, with its eigenvalue of 51.2. The first contrast (the largest secondary dimension) had an eigenvalue of 1.7 for 2.9% of the unexplained variance. The fit statistics for the Mathematics items were mostly within acceptable limits: the infit MNSQ ranged from 0.71 to 1.75; the outfit MNSQ ranged from 0.72 to 1.57. Item 20.C (connects numerals with their quantities) yielded MNSQ statistics that were beyond the acceptable range (1.75 and 1.57). This item did, however, yield an item total score correlation of .89, illustrating that it does provide information that is related to the rest of the information provided by this set of items. The item total score correlations ranged from .89 to .95.

Physical Scale (5 items)

The PCAR showed that for the Physical scale, the Rasch dimension explained 89.9% of the variance in the data, with its eigenvalue of 44.6. The first contrast (the largest secondary dimension) had an eigenvalue of 1.7 and accounted only for 3.4% of the unexplained variance. The fit statistics for all of the Physical items were mostly within acceptable limits: the infit MNSQ ranged from 0.79 to 1.39; the outfit MNSQ ranged from 0.82 to 1.45. Item 7.B (uses writing and drawing tools) yielded MNSQ statistics that were close to or slightly beyond the acceptable range (1.39 and 1.45). This item did, however, yield an item total score correlation of .94, illustrating that it does provide information that is related to the rest of the information provided by this set of items. All five of the item total score correlations were .94.

Social Emotional Scale (9 items)

The PCAR showed that for the Social Emotional scale, the Rasch dimension explained 87.7% of the variance in the data, with its eigenvalue of 64.4. The first contrast (the largest secondary dimension) had an eigenvalue of 2.4 and accounted only for 3.3% of the unexplained variance. The fit statistics for all of the Social Emotional items were well within acceptable limits:

the infit MNSQ ranged from 0.76 to 1.30; the outfit MNSQ ranged from 0.76 to 1.28. The item total score correlations ranged from .89 to .94.

In summary, with a 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. The exceptions to this conclusion where the results suggest the possibility of item misfit within a given scale need to be monitored and evaluated again in the future as teachers across the country gain more experience using the GOLD assessment system.

Rating Category Effectiveness

The items are measured on a 10-point scale labeled 0 through 9. The use of rating scale categories was examined, which can provide information about whether teachers utilize the instrument in the manner in which it was intended. It is recommended that for each item, each rating scale category is assigned to a minimum of 10 children. The average of the ability estimates for all persons in the sample who chose that particular response category was examined (Bond & Fox, 2007). Average measure scores should advance monotonically with rating scale category values. Thresholds (also called step calibrations) are the difficulties estimated for choosing one response category over another (Bond & Fox, 2007). Thresholds should also increase monotonically with rating scale category. The magnitudes 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 probability curve graph (Bond & Fox, 2007).

For all of the items with 10 point rating scales, the teachers used all 10 rating scale points and there were sufficient observations in each of the categories to model the ratings scale. Items 19.A (writes name) and 19.B (writes to convey meaning) offer teachers only an 8 point rating scale and all 8 points were used. For all six scales, the average measure increased with the category level

and the thresholds advanced with the categories. An examination of the Rasch category probability curves indicated that all of the categories were distinct. In general, the pattern was very similar across all the scales.

Item Difficulty Measures

For all six scales, the item location hierarchy appeared to be generally consistent with the expected developmental trajectory for typically developing children.

For the Cognitive Scale, the item pertaining to a child's use of classification skills (13) was found to be the most difficult item. The items pertaining to a child's ability to attend and engage (11.A) and show motivation and interest (11.D) were estimated as the easiest items. The range of overall item difficulties (-1.39 to 1.64) item anchor point locations was considered sufficient for separation of children across the range of underlying abilities.

For the Language Scale, the item pertaining to a child's ability to describe another place or time (9.D) was found to be the most difficult item. The item pertaining to a child's ability to speak clearly (9.B) was estimated as the easiest item. The range of item difficulties (-1.15 to 2.64) and item anchor point locations was considered wide enough for reasonable separation of children according to underlying ability.

For the Literacy Scale, the item pertaining to a child's use of letter-sound knowledge (16.B) was found to be the most difficult item. The item pertaining to a child's knowledge of print (17A) was estimated as the easiest item. The range of both item difficulties (-2.01 to 1.55) and item anchor point locations 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 connect numerals to quantities (20.C) was found to be the most difficult item. The item pertaining to a child's ability to explore shapes (21.A) was estimated to be the easiest item. The range of both item difficulties (-.81

to 1.19) and item anchor point locations, although narrower that for the other scales and based on somewhat fewer items, 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 use writing and drawing tools (7.B) was found to be the most difficult item. The item pertaining to a child's ability to demonstrate balancing skills (5) was estimated as the easiest item. The range of overall item difficulties (-.39 to 1.74) and item anchor point locations, although narrower that for the other scales and based on somewhat fewer items, was considered wide enough for reasonable separation of children according to underlying ability.

For the Social Emotional Scale, the item pertaining to a child's ability to balance the needs and rights of self and others (3.A) was found to be the most difficult item. The item pertaining to a child's ability to form relationships with adults (2.A) was estimated as the easiest item. The range of both item difficulties (-2.26 to 1.38) and item anchor point locations 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 developmental theory. In addition, the range of difficulties for each scale is the widest that has been observed with data from our norm samples to date, suggesting that teachers in the field are getting much better at separating children according to underlying ability and performance as they gain more experience with the use of the assessment. It is also important to recognize that the range of item difficulties is effectively much wider than these results indicate when considering the separation created between children by the range of rating scale anchor point threshold locations.

Reliability

Reliability was evaluated using Cronbach's alpha measure of internal consistency, and the person separation index, item separation index, person reliability, and item reliability provided by Winsteps. 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 ideal (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 by Winsteps is equivalent 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).

Cognitive Scale

Based on the Rasch reliability indexes (see Table 3), the scale scores appear to be highly reliable, as evidenced by person separation indexes of 7.69, person reliabilities of .98, item separation indexes of 95.28, and item reliabilities of .99. The Cronbach's alpha reliability coefficient for this scale was .99, indicating high internal consistency reliability.

Language Scale

Based on the Rasch reliability indexes, the scale scores appear to be highly reliable, as evidenced by person separation indexes of 6.96, person reliabilities of .98, item separation indexes of 119.41, and item reliabilities of .99. The Cronbach's alpha reliability coefficient for this scale was .98, indicating high internal consistency reliability.

Literacy Scale

Based on the Rasch reliability indexes, the scale scores appear to be highly reliable, as evidenced by person separation indexes of 4.89, person reliabilities of .96, item separation indexes of 88.06, and item reliabilities of .99. The Cronbach's alpha reliability coefficient for this scale was .98, indicating high internal consistency reliability.

Mathematics Scale

Based on the Rasch reliability indexes, the scale scores appear to be reliable, as evidenced by person separation indexes of 4.61, person reliabilities of .96, item separation indexes of 66.15, and item reliabilities of 0.99. The Cronbach's alpha reliability coefficient for this scale was .98, indicating high internal consistency reliability.

Physical Scale

Based on the Rasch reliability indexes, the scale scores appear to be highly reliable, as evidenced by person separation indexes of 4.93, person reliabilities of .96, item separation indexes of 84.18, and item reliabilities of .99. The Cronbach's alpha reliability coefficient for this scale was .97, indicating acceptable internal consistency reliability.

Social Emotional Scale

Based on the Rasch reliability indexes, the scale appear to be highly reliable, as evidenced by person separation indexes of 6.11, person reliabilities of .97, item separation indexes of 133.36, and item reliabilities of .99. The Cronbach's alpha reliability coefficient for this scale was .98, indicating high internal consistency reliability.

Scale Scores, Raw Scores, and Norm Tables

For the purpose of creating norm tables for the GOLD assessment system, fall, winter, and spring assessment data were used. The scale scores were created by first calculating raw scores for each child. If a child did not have complete rating data, but was rated by the teacher on at least 80% of the items on a respective scale, then the child's scale mean rating was substituted for the missing ratings. The scale scores were created by transforming the raw scores into interval level Rasch rating scale ability estimates for each child. The ability estimates were then scaled to conform to a distribution with a mean of 500 and standard deviation of 100. The winter data was used to calibrate the scaling.

The raw score to scale conversion tables generated by the Rasch PCM, based on the national norm data, were used to rescale the raw scores into scale scores. Scale scores values 3 or more standard deviations below the mean were given a value of 200 and values three or more standard deviations above the mean were given a value of 800. This scaling strategy is commonly used in educational and psychological testing.

For each scale score and age / grade band, as shown in Tables 4 - 9, 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 and age / grade band. In all IRT models, unlike with classical measurement models, the SEM can be estimated for each scale score point. Tables 10 - 15 contain similar statistics for the raw scores. These raw scores consist of summing the item ratings for each scale. The mean, standard deviation, and quartile boundaries (25^{th}, 50^{th}, 75^{th} percentiles) were calculated for the distribution of raw scores for all three time points and for fall to spring gains. The SEM values were calculated using classical measurement theory.

Summary

Overall, the GOLD assessment system 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. The results 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 the GOLD assessment system 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 strong 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 the GOLD assessment can be used by teachers to help them understand the developmental trajectory that most children will follow.

Future research using data from this particular source 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.

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.

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
Norm sample by ethnic group

Group	2013 Census Bureau Estimates		2013-14 GOLD Norm Sample	
White	12,437,432	51.3%	15,484	51.6%
African American	3,325,732	13.7%	4,110	13.7%
Native American	211,371	0.9%	264	0.9%
Asian	1,077,907	4.4%	1,231	4.1%
Hawaiian / Pacific Islander	47,064	0.2%	60	0.2%
Multiple Races	1,057,269	4.4%	1,303	4.3%
Hispanic	6,101,445	25.2%	7,548	25.2%
Total	24,258,220	100.0%	30,000	100.0%

Table 2
Norm sample by child characteristics

Child Characteristic	Levels	Count	Percentage
Color Band	Infants	5,000	16.7%
	One year olds	5,000	16.7%
	Two year olds	5,000	16.7%
	Prekindergarten 3	5,000	16.7%
	Prekindergarten 4	5,000	16.7%
	Kindergarten	5,000	16.7%
Gender	Male	15,328	51.2%
	Female	14,592	48.8%
Disability Status	IFSP	1,197	4.0%
	IEP	1,707	5.7%
	IFSP and / or IEP	2,850	9.5%
Lunch Status	Free or Reduced	8,218	27.4%
	Pay	21,782	72.6%
Primary Language Spoken in the Home	English	23,809	79.4%
	Spanish	4,466	14.9%
	Other	1,725	5.8%

Table 3
Reliability statistics by scale

Scale	Cronbach's Alpha	Item Reliability	Item Separation Index	Person Reliability	Person Separation Index
Cognitive	0.99	0.99	95.28	0.98	7.69
Language	0.98	0.99	119.41	0.98	6.96
Literacy	0.98	0.99	88.06	0.96	4.89
Mathematics	0.98	0.99	66.15	0.96	4.61
Physical	0.97	0.99	84.18	0.96	4.93
Social Emotional	0.98	0.99	133.36	0.97	6.11

Table 4
Cognitive standard scores by checkpoint and color band

Colorband		Fall	Winter	Spring	Fall to Spring Growth
Infants	Mean	356.66	379.70	401.63	45.09
	SD	29.56	32.04	35.15	28.34
	25th percentile	339	362	376	27
	50th percentile	357	376	399	42
	75th percentile	371	399	423	61
	SEM	10	10	10	
One year olds	Mean	419.85	442.84	464.95	45.21
	SD	37.45	40.48	46.11	36.48
	25th percentile	394	418	438	21
	50th percentile	418	438	459	42
	75th percentile	444	464	489	64
	SEM	10	10	10	
Two year olds	Mean	460.94	487.18	512.47	52.37
	SD	44.46	47.17	54.73	45.09
	25th percentile	433	459	479	26
	50th percentile	459	484	507	48
	75th percentile	484	513	538	72
	SEM	10	11	11	
Prekindergarten 3	Mean	508.59	550.42	588.02	80.59
	SD	58.53	61.05	68.61	52.41
	25th percentile	474	513	544	49
	50th percentile	507	550	584	74
	75th percentile	538	584	626	107
	SEM	11	11	11	
Prekindergarten 4	Mean	559.11	611.34	660.45	102.02
	SD	59.85	60.88	69.66	58.49
	25th percentile	525	573	619	63
	50th percentile	562	613	655	96
	75th percentile	595	647	704	134
	SEM	11	11	12	
Kindergarten	Mean	617.74	679.85	730.39	109.88
	SD	68.48	75.07	68.92	52.52
	25th percentile	579	640	697	72
	50th percentile	626	691	745	109
	75th percentile	662	724	796	145
	SEM	12	12	12	

Table 5
Language standard scores by checkpoint and color band

Colorband		Fall	Winter	Spring	Fall to Spring Growth
Infants	Mean	364.66	386.43	406.07	41.61
	SD	29.60	29.82	30.00	26.12
	25th percentile	349	368	389	25
	50th percentile	368	389	403	40
	75th percentile	384	403	426	55
	SEM	11	10	10	
One year olds	Mean	422.83	444.42	465.83	43.16
	SD	35.01	39.04	45.33	33.49
	25th percentile	403	421	439	23
	50th percentile	421	443	462	40
	75th percentile	443	467	491	59
	SEM	9	10	11	
Two year olds	Mean	465.76	489.47	515.62	50.64
	SD	45.18	49.25	59.13	44.32
	25th percentile	439	457	479	24
	50th percentile	462	485	509	46
	75th percentile	491	514	539	69
	SEM	11	11	11	
Prekindergarten 3	Mean	508.72	547.24	585.92	77.95
	SD	63.15	67.87	78.02	55.37
	25th percentile	467	503	539	43
	50th percentile	509	545	585	71
	75th percentile	545	585	633	104
	SEM	11	12	14	
Prekindergarten 4	Mean	560.35	611.17	663.92	104.17
	SD	66.40	70.95	80.12	62.96
	25th percentile	520	567	614	64
	50th percentile	560	614	670	100
	75th percentile	595	651	718	141
	SEM	12	14	14	
Kindergarten	Mean	612.22	677.20	726.39	108.00
	SD	78.89	83.05	77.72	57.29
	25th percentile	560	633	689	70
	50th percentile	614	680	745	108
	75th percentile	670	730	797	146
	SEM	14	14	16	

Table 6
Literacy standard scores by checkpoint and color band

Colorband		Fall	Winter	Spring	Fall to Spring Growth
Infants	Mean	348.47	373.61	397.44	49.19
	SD	37.65	41.18	40.96	38.66
	25th percentile	306	348	374	26
	50th percentile	348	374	402	44
	75th percentile	374	402	421	72
	SEM	35	26	19	
One year olds	Mean	420.20	441.45	460.14	40.72
	SD	38.61	39.90	40.65	35.05
	25th percentile	402	421	435	19
	50th percentile	421	435	457	38
	75th percentile	441	467	484	59
	SEM	16	14	13	
Two year olds	Mean	459.81	480.98	499.90	41.12
	SD	39.51	41.31	45.26	36.30
	25th percentile	441	457	476	19
	50th percentile	462	480	497	36
	75th percentile	480	501	521	58
	SEM	13	12	12	
Prekindergarten 3	Mean	514.94	551.65	581.37	67.99
	SD	50.69	52.56	58.56	40.50
	25th percentile	484	517	544	42
	50th percentile	513	547	576	64
	75th percentile	544	579	614	90
	SEM	12	11	11	
Prekindergarten 4	Mean	560.04	606.50	645.80	86.07
	SD	51.48	52.50	57.56	43.12
	25th percentile	529	572	610	57
	50th percentile	558	607	648	82
	75th percentile	593	638	683	111
	SEM	11	11	11	
Kindergarten	Mean	623.13	698.72	738.49	103.17
	SD	61.73	56.91	48.51	47.91
	25th percentile	583	669	720	72
	50th percentile	628	705	751	101
	75th percentile	665	741	763	132
	SEM	11	11	12	

Table 7
Mathematics standard scores by checkpoint and color band

Colorband		Fall	Winter	Spring	Fall to Spring Growth
Infants	Mean	344.05	360.71	385.05	41.66
	SD	23.17	37.86	46.07	42.62
	25th percentile	336	336	336	0
	50th percentile	336	336	387	31
	75th percentile	336	387	428	84
	SEM	44	25	19	
One year olds	Mean	417.07	442.25	464.13	47.49
	SD	43.30	42.73	42.09	37.33
	25th percentile	387	428	443	22
	50th percentile	428	443	465	43
	75th percentile	443	465	486	69
	SEM	14	13	13	
Two year olds	Mean	464.99	487.39	508.36	45.00
	SD	40.99	39.83	42.52	35.60
	25th percentile	443	465	486	22
	50th percentile	465	486	511	41
	75th percentile	486	511	533	62
	SEM	13	13	12	
Prekindergarten 3	Mean	517.22	550.90	579.52	63.31
	SD	49.37	49.49	55.69	40.65
	25th percentile	486	522	549	38
	50th percentile	522	549	577	59
	75th percentile	544	577	606	84
	SEM	12	11	12	
Prekindergarten 4	Mean	560.00	601.85	639.53	80.17
	SD	48.85	49.63	56.68	43.30
	25th percentile	533	572	606	52
	50th percentile	560	600	637	76
	75th percentile	589	631	671	103
	SEM	12	12	13	
Kindergarten	Mean	616.31	679.06	726.73	103.42
	SD	55.37	57.04	53.40	46.18
	25th percentile	583	644	698	72
	50th percentile	618	685	741	102
	75th percentile	651	713	771	134
	SEM	12	13	15	

Table 8
Physical standard scores by checkpoint and color band

Colorband		Fall	Winter	Spring	Fall to Spring Growth
Infants	Mean	347.60	377.68	405.15	57.92
	SD	38.74	41.97	46.74	37.70
	25th percentile	327	350	375	34
	50th percentile	350	375	399	56
	75th percentile	367	399	435	75
	SEM	14	13	14	
One year olds	Mean	427.07	450.16	472.36	45.37
	SD	46.82	49.11	54.58	46.20
	25th percentile	399	425	435	21
	50th percentile	425	447	469	44
	75th percentile	447	480	501	67
	SEM	15	16	16	
Two year olds	Mean	467.54	492.82	518.64	51.67
	SD	53.80	53.49	60.03	52.53
	25th percentile	435	458	491	22
	50th percentile	458	491	512	47
	75th percentile	501	522	546	76
	SEM	16	15	15	
Prekindergarten 3	Mean	506.22	545.90	583.47	77.95
	SD	62.97	63.29	68.49	56.78
	25th percentile	469	512	546	41
	50th percentile	512	546	573	68
	75th percentile	546	587	625	106
	SEM	15	17	17	
Prekindergarten 4	Mean	557.27	607.33	658.50	101.35
	SD	61.35	59.52	67.20	62.93
	25th percentile	522	573	613	61
	50th percentile	559	613	663	92
	75th percentile	587	638	701	138
	SEM	17	17	18	
Kindergarten	Mean	618.73	672.34	707.89	88.94
	SD	63.68	65.93	63.04	58.13
	25th percentile	587	638	679	51
	50th percentile	625	663	740	90
	75th percentile	663	740	740	126
	SEM	17	20	25	

Table 9
Social Emotional standard scores by checkpoint and color band

Colorband		Fall	Winter	Spring	Fall to Spring Growth
Infants	Mean	352.48	381.86	407.57	55.45
	SD	39.21	40.06	42.72	36.32
	25th percentile	334	355	379	33
	50th percentile	355	379	402	52
	75th percentile	373	402	430	74
	SEM	14	13	13	
One year olds	Mean	423.76	448.35	470.27	46.44
	SD	43.50	44.46	48.29	41.61
	25th percentile	396	419	442	23
	50th percentile	425	448	465	44
	75th percentile	448	470	497	67
	SEM	13	13	13	
Two year olds	Mean	461.79	486.88	511.61	50.68
	SD	49.11	49.85	56.39	48.29
	25th percentile	436	459	481	22
	50th percentile	459	486	508	44
	75th percentile	486	508	531	72
	SEM	13	13	13	
Prekindergarten 3	Mean	502.94	545.77	583.98	81.62
	SD	61.86	62.53	70.56	54.73
	25th percentile	465	508	543	47
	50th percentile	503	543	579	75
	75th percentile	537	579	623	109
	SEM	13	13	13	
Prekindergarten 4	Mean	555.25	607.54	657.96	103.09
	SD	62.87	63.65	72.22	63.12
	25th percentile	520	573	610	61
	50th percentile	561	610	652	96
	75th percentile	591	645	700	136
	SEM	13	14	15	
Kindergarten	Mean	616.64	679.92	721.50	102.99
	SD	72.49	75.98	73.33	63.28
	25th percentile	573	630	675	67
	50th percentile	623	683	731	103
	75th percentile	660	731	786	146
	SEM	14	15	18	