CEMETR-2017-02
MARCH-2017

CEME
Technical Report

The Center for Educational Measurement and Evaluation

Technical Manual for the Teaching Strategies GOLD™ Assessment System: Birth through Third Grade Edition

Richard G. Lambert

RICHARD LAMBERT
CHUANG WANG
MARK D'AMICO
SERIES EDITORS

A PUBLICATION OF
THE CENTER FOR
EDUCATIONAL
MEASUREMENT
AND EVALUATION

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

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

Birth through Third Grade Edition

Richard G. Lambert, Ph.D.

Center for Educational Measurement and Evaluation

UNC Charlotte

March, 2017

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

The GOLD measure yields information that is rooted in the ongoing work of teachers as they develop and collect evidences that are used to identify the best fits for each child across a series of developmental progressions. Teachers collect ongoing portfolios of evidences throughout the academic year, reflect upon and analyze those evidences, make preliminary ratings on an ongoing basis, and finalize ratings at specified points during the year. This information is intended to be used to inform instruction and to facilitate communication with parents and other stakeholders. In contrast to direct assessments, evidences are collected within regular activities in natural classroom contexts. 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 birth to third grade.

Background Information on the Development of GOLD

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

The development of GOLD occurred over several years and incorporated feedback from teachers, administrators, consultants, and Teaching Strategies, LLC professional-development and research personnel. Pilot studies with diverse populations were conducted, and a draft of the measure was sent to leading authorities in the field for content review. Major revisions were made based on results of the content validation and pilot studies. Final assessment items were selected on the basis of feedback received during the development process; state early learning standards and the *Head Start Child Development and Early Learning Framework* (U.S. Department of Health & Human Services, 2010); and current research and professional literature including literature that identifies which knowledge, skills, and 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 GOLD measure has been expended to include more rating scale items and additional rating scale categories in order to incorporate developmental expectations for children up to third grade. The 23 GOLD objectives included in the current studies are now operationalized into 60 rating scale items: social–emotional (9 items), physical (5 items), language (8 items), cognitive (10 items), literacy (16 items) and mathematics (12 items). Teachers rate children's skills, knowledge, and behaviors along rating scales that range from 10 to 19 points and outline progressions of development and learning. These progressions range from "Not Yet" (Level 0) to "Exceeds Third Grade Expectations" (Levels 9 to 19 depending upon the progression). Each progression includes indicator levels with varied examples from everyday situations that give teachers guidance of what evidence may look like. There are also "In-between" levels and do not include examples. They allow for additional steps in the progression as the child demonstrates that skills are emerging in a particular area but are not fully established. Overlapping, color-coded bands indicate the typical age and/or grade-level (i.e., kindergarten) ranges for each item measured.

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 nine 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, 6.) kindergarten, 7.) first grade, 8.) second grade, and 9.) third grade. After reducing the population to cases that met these criteria, stratified random sampling, stratifying on ethnicity and age, was used to select 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 2015 Census Bureau national estimates for the proportion of children ages birth to 17 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 33,294 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 closely replicated the national Census Bureau 2015 estimates. White children who were slightly over represented and African American, Asian, and Hispanic children who were slightly under represented.

As shown in Table 2, the norm sample was evenly balanced by gender (boys=51.2%, girls=48.8%). Children with an IEP or IFSP comprised 9.8% of the norm sample. A total of 26.4% of the norm sample qualified for free or reduced price lunch. The primary language spoken in the home was distributed as follows: English (80.0%), Spanish (13.8%), and other languages (6.2%).

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 have different rating scale structures (i.e., number of rating scale categories and labels across items). Specifically, 12 items include a 0-9 scale, 6 items include a 0-11 scale, 16 items include a 0-13 scale, 25 items include a 0-15 scale, and one item includes a 0-19 scale. For each item, the 0 category represents "Not Yet" and the highest category represents abilities beyond the highest behavioral anchor. 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). Infit MNSQ statistics represent the fit of individual item response patterns to the measurement model, address possible secondary dimensions, and indicate fit to the underlying construct. Outfit MNSQ statistics are sensitive to outliers, that is response patterns 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, 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 90.5% of the variance in the data, with its eigenvalue of 105.2. The first contrast (the largest secondary dimension) had an eigenvalue of 2.3 and accounted for only 2.2% of the unexplained variance. The fit statistics for all of the Cognitive items were within acceptable limits: the infit MNSQ ranged from 0.79 to 1.19; the outfit MNSQ ranged from 0.80 to 1.21. The item total score correlations ranged from .94 to .96.

Language Scale (8 items)

The PCAR showed that for the Language scale, the Rasch dimension explained 91.7% of the variance in the data, with its eigenvalue of 96.3. The first contrast (the largest secondary dimension) had an eigenvalue of 1.9 and accounted for only 2.0% of the unexplained variance. The fit statistics for all of the Language items were well within acceptable limits: the infit MNSQ ranged from 0.76 to

1.14; the outfit MNSQ ranged from 0.81 to 1.17. The item total score correlations ranged from .93 to .95.

Literacy Scale (16 items)

The PCAR showed that the Rasch dimension explained 85.7% of the variance in the data, with its eigenvalue of 111.7. The first contrast (the largest secondary dimension) had an eigenvalue of 3.4 and accounted for only 3.0% of the unexplained variance. The item total score correlations ranged from .57 to .95. The fit statistics for the Literacy items were mostly within acceptable limits: the infit MNSQ ranged from 0.61 to 2.30; the outfit MNSQ ranged from 0.56 to 1.96. Two items showed fit statistics outside the acceptable range. Item 17.A (uses and appreciates books and other text) yielded MNSQ statistics that were beyond the acceptable range (1.88 and 2.12). This item did, however, yield an item total score correlation of .92, illustrating that it does provide information that is related to the rest of the information provided by this set of items. Item 19.A (writes name) yielded MNSQ statistics that were beyond the acceptable range (2.30 and 1.96). This item did, however, yield an item total score correlation of .87, illustrating that it does provide information that is related to the rest of the information provided by this set of items. 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). Further research will need to focus on the fit statistics for these and any items with potential fit issues.

Mathematics Scale (12 items)

The PCAR showed that the Rasch dimension explained 91.70% of the variance in the data, with its eigenvalue of 145.2. The first contrast (the largest secondary dimension) had an eigenvalue of 2.1 for only 1.4% of the unexplained variance. The fit statistics for the Mathematics items were mostly within acceptable limits: the infit MNSQ ranged from 0.70 to 1.84; the outfit MNSQ ranged from 0.37 to 2.31. The item total score correlations ranged from .55 to .95. Three items showed fit statistics outside the acceptable range. Item 20.D (understands and uses place value and base 10) yielded MNSQ statistics that were beyond the acceptable range (.86 and .37). This item did, however, yield an item total score correlation of .55, illustrating that it does provide some information that is related to the rest of the information provided by this set of items. Item 20.E (applies properties of mathematical operations and relationships) yielded MNSQ statistics that were beyond the acceptable range (.78 and .1.96). This item did, however, yield an item total score correlation of .63, illustrating that it does provide some information that is related to the rest of the information provided by this set of items. Item 20.F (applies number combinations and mental number strategies) yielded MNSQ statistics that were beyond the acceptable range (1.21 and 2.31). This item did, however, yield an item total score correlation of .59, illustrating that it does provide some information that is related to the rest of the information provided by this set of items.

Physical Scale (5 items)

The PCAR showed that for the Physical scale, the Rasch dimension explained 90.4% of the variance in the data, with its eigenvalue of 51.8. The first contrast (the largest secondary dimension) had an eigenvalue of 1.7 and accounted for only 3.3% 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 1.36; the outfit MNSQ ranged from 0.81 to 1.42. Item 7.B (uses writing and drawing tools) yielded MNSQ statistics that were close to or slightly beyond the acceptable range (1.36 and 1.42). This item did, however, yield an item total score correlation of .95, 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 .95.

Social Emotional Scale (9 items)

The PCAR showed that for the Social Emotional scale, the Rasch dimension explained 88.6% of the variance in the data, with its eigenvalue of 79.1. The first contrast (the largest secondary dimension) had an eigenvalue of 2.2 and accounted for only 2.8% of the unexplained variance. The fit statistics for all of the Social Emotional items were well within acceptable limits: the infit MNSQ ranged from 0.71 to 1.34; the outfit MNSQ ranged from 0.72 to 1.35. The item total score correlations ranged from .90 to .95.

In summary, with the few exceptions noted above, these model fit statistics when taken together generally suggest that the data does in fact fit the Rasch PCM very well. These results also indicated that the data satisfied the unidimensionality assumption of the Rasch model. 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 B-3 assessment system.

Rating Category Effectiveness

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 six scales, the average measure score 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. However, there were a number of issues related to the use of all of the categories by the teachers making the ratings of the children. These results may have been impacted by the relatively small number of children in the norm sample representing the older age grade categories (first through third grades).

This issue can be seen in the results for all of the scales. For the Cognitive scale, teachers did not use all of the rating scale categories for 7 of the 10 items with the highest category not used. In addition, for 9 of the 10 items there were not at least 10 children placed in all of the remaining categories. For the Language scale, teachers did not use all of the rating scale categories for 7 of the 8 items with the highest category not used. In addition, for 7 of the 8 items there were not at least 10 children placed in all of the remaining categories. For the Literacy scale, teachers did not use all of the rating scale categories for 9 of the 16 items with the highest category not used. In addition, for 2 of the 16 items there were not at least 10 children placed in all of the remaining categories. For the Mathematics scale, teachers did not use all of the rating scale categories for 11 of the 12 items with the highest category not used. In addition, for 9 of the 12 items there were not at least 10 children placed in all of the remaining categories. For the Physical scale, teachers did not use all of the rating scale categories for all 5 of the items with the highest category not used. In addition, for 1 of the 5 items there were not at least 10 children placed in all of the remaining categories. For the Social Emotional scale, teachers did not use all of the rating scale categories for 6 of the 9 items with the highest category not used. In addition, for 7 of the 9 items there were not at least 10 children placed in all of the remaining categories. These issues need to be monitored in future research and underscore the importance of a more extensive norm sample for the older age grade groups. These results may also suggest the need for further teacher training related to the meaning and use of all of the rating scale categories and behavioral anchors.

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 item pertaining to a child's ability to attend and engage (11.A) was estimated as the easiest items. The range of overall item difficulties (-1.07 to 1.28) and 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 (-.87 to 2.16) 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 using context clues to read and comprehend text (18.D) was found to be the most difficult item. The item pertaining to name writing (19.A) and using and appreciating books and print were estimated as the easiest items. The range of both item difficulties (-2.22 to 3.63) 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 understanding and using place values and base 10 (20.D) was found to be the most difficult item. The item pertaining to a child's ability to count (20.A) was estimated to be the easiest item. The range of both item difficulties (-3.26 to 5.39) and item anchor point locations 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 traveling skills (4) was estimated as the easiest item. The range of overall item difficulties (-.53 to 1.33) 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 (-1.91 to 1.11) 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 7.81, person reliabilities of .98, item separation indexes of 85.38, 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 7.01, person reliabilities of .98, item separation indexes of 110.60, and item reliabilities of .99. The Cronbach's alpha reliability coefficient for this scale was .99, 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 3.51, person reliabilities of .93, item separation indexes of 226.01, and item reliabilities of .99. The Cronbach's alpha reliability coefficient for this scale was .97, 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 4.02, person reliabilities of .94, item separation indexes of 278.17, and item reliabilities of 0.99. The Cronbach's alpha reliability coefficient for this scale was .96, 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 5.28, person reliabilities of .97, item separation indexes of 76.62, 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 scores appear to be highly reliable, as evidenced by person separation indexes of 6.32, person reliabilities of .98, item separation indexes of 132.50, and item reliabilities of .99. The Cronbach's alpha reliability coefficient for this scale was .98, indicating high internal consistency reliability.

Scale Scores

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. The scale scores have been created to have a range of 0 to 1,000. For each scale score and age / grade band, as shown in Tables 4, the scale mean, standard deviation, and quartile boundaries (25th, 50th, 75th percentiles) are reported for the winter checkpoint. The standard errors of measurement (SEM) are reported at the scale mean for each scale and age / grade band. In all IRT models, unlike with classical measurement models, the SEM can be estimated for each scale score point.

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 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 some 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.

However, future research will need to focus on whether teachers can use all of the categories in the rating scales. It will be important to develop norm samples that include larger samples of children in the older age grade groups. Future research will also be needed to focus on the degree of association between GOLD B-3 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 B-3 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.

Lambert, R., Kim, D., & Burts, D. (2015). The measurement properties of the Teaching Strategies GOLD assessment system. *Early Childhood Research Quarterly*, *33, 49-63*.

Lambert, R. & Kim, D-H., & Burts, D. (2014). *Technical manual for the Teaching Strategies GOLD assessment system (3rd edition)*. 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 ethnic group

Group	2015 Census Bureau Estimates	GOLD B-3 Norm Sample	
White	51.5%	18,362	55.2%
African American	13.8%	4,175	12.5%
Native American	0.9%	268	0.8%
Asian	4.9%	1,294	3.9%
Hawaiian / Pacific Islander	0.2%	107	0.3%
Multiple Races	4.1%	1,450	4.4%
Hispanic	24.6%	7,638	22.9%
Total	100.0%	33,294	100.0%

Table 2
Norm sample by child characteristics

Child Characteristic	Levels	Count	Percentage
Color Band	Infants	5,000	15.0%
	One year olds	5,000	15.0%
	Two year olds	5,000	15.0%
	Prekindergarten 3	5,000	15.0%
	Prekindergarten 4	7,734	23.2%
	Kindergarten	5,000	15.0%
	First grade	314	0.9%
	Second grade	191	0.6%
	Third grade	55	0.2%
Gender	Male	17,003	51.2%
	Female	16,220	48.8%
Disability Status	Yes	30,026	90.2%
	No	3,268	9.8%
Lunch Status	Free or Reduced	8,792	26.4%
	Pay	24,502	73.6%
Primary Language Spoken in the Home	English	24,931	80.0%
	Spanish	4,295	13.8%
	Other	1,928	6.2%

Table 3
Reliability indexes by scale

	Items	Number of Items	Person Separation Index	Person Reliability	Cronbach's Alpha	Item Separation Index	Item Reliability
Social Emotional	1a - 3b	9	6.32	0.98	0.98	132.50	0.99
Physical	4 - 7b	5	5.28	0.97	0.97	76.62	0.99
Language	8a - 10b	8	7.01	0.98	0.99	110.60	0.99
Cognitive	11a - 14b	10	7.81	0.98	0.99	85.38	0.99
Literacy	15a - 19c	16	3.51	0.93	0.97	226.01	0.99
Mathematics	20a - 23	12	4.02	0.94	0.96	278.17	0.99

Table 4

Winter scale scores by age / grade

							Social
Age / Grade		Cognitive	Language	Literacy	Mathematics	Physical	Emotional
Birth to 1 year	Mean	160.21	140.16	153.01	42.61	198.06	183.28
	SD	47.99	40.41	89.24	62.81	66.39	42.52
	25th	133	115	100	0	155	155
	50th	155	144	157	0	193	180
	75th	181	163	217	94	231	204
	SEM	15	15	59	47	23	15
1 to 2 years	Mean	254.46	217.79	293.21	166.43	313.13	253.72
	SD	60.13	52.15	78.80	58.21	78.36	46.76
	25th	217	187	255	154	273	223
	50th	248	216	282	172	307	254
	75th	286	249	344	197	361	277
	SEM	16	14	31	21	27	15
2 to 3 years	Mean	320.23	277.71	369.16	221.99	381.40	294.20
	SD	69.85	65.45	77.34	49.65	85.22	51.94
	25th	278	235	326	197	325	265
	50th	316	271	370	220	378	294
	75th	358	311	408	249	429	317
	SEM	17	16	26	19	26	14
Preschool 3	Mean	413.58	353.94	493.86	301.86	465.91	355.16
	SD	89.41	88.49	87.38	63.26	98.70	63.97
	25th	358	295	438	263	412	317
	50th	414	353	492	299	467	353
	75th	464	405	547	338	533	390
	SEM	18	17	23	19	29	15
Prekindergarten 4	Mean	497.54	439.67	639.08	429.80	568.74	418.96
	SD	93.39	91.32	82.54	76.38	97.48	67.60
	25th	448	393	601	392	511	383
	50th	497	443	651	438	575	422
	75th	546	503	696	474	634	457
	SEM	18	20	19	18	28	15
Kindergarten	Mean	585.32	520.66	727.61	499.88	669.09	476.95
	SD	103.02	97.32	65.22	67.03	112.90	71.08
	25th	546	491	705	474	614	442
	50th	600	526	739	506	677	479
	75th	646	574	763	540	727	519
	SEM	19	19	15	16	30	16
1st grade	Mean	748.63	657.72	808.96	632.20	799.46	599.83
	SD	106.70	94.26	57.43	66.40	70.21	90.03
	25th	743	601	800	607	793	556
	50th	789	694	821	646	806	624
	75th	804	717	832	670	840	661
	SEM	15	19	12	16	22	18
2nd grade	Mean	758.59	692.75	823.64	678.29	843.35	649.20
	SD	160.09	153.87	69.91	110.17	104.28	120.67
	25th	717	601	811	670	840	604
	50th	838	755	835	724	863	682
	75th	863	806	861	739	893	727
	SEM	15	16	12	15	22	16
3rd grade	Mean	788.64	709.57	874.07	800.36	889.86	670.57
	SD	104.50	131.03	34.49	78.60	41.20	65.57
	25th	799	716	866	809	883	656
	50th	813	740	876	817	893	689
	75th	838	757	904	822	904	697
	SEM	13	16	14	16	20	16