CEME

Technical Report

CEMETR-2015-04
NOVEMBER-2015

The Center for Educational Measurement and Evaluation

Using Teaching Strategies Gold to Assess Kindergarten Readiness and Track Growth and Development

Richard Lambert
Do-Hong Kim
Diane Burts

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

Using Teaching Strategies Gold to Assess Kindergarten Readiness and Track Growth and Development

Richard Lambert, Do-Hong Kim, and Diane Burts

Center for Educational Measurement and Evaluation

The University of North Carolina at Charlotte

November, 2015

This report is focused on an evaluation of the measurement properties of the scale scores and teacher ratings that result from the use of the GOLD assessment system with children in kindergarten classrooms. Several states have chosen to implement GOLD widely for kindergarten entry assessment. In many other states GOLD is one of a variety of formative assessments available to teachers as they gather evidence to understand the kindergarten readiness of the children entering their classrooms. In still other settings, teachers are using GOLD to track the growth and development of the children in their classrooms across the academic year. The purpose of this report is to examine statistical indexes of reliability and validity based on the information produced using the GOLD formative assessment system for both kindergarten entry and across the kindergarten academic year.

Sample

A total of 123,980 kindergarten children had skills rated using the GOLD assessment system at some point during the 2014-2015 academic year. These children were 51.5% male and 48.5% female. Typically developing children comprised 96.3% of the sample. Children with an IEP comprised the remaining 3.7% of the sample. Children from economically disadvantaged backgrounds who qualify for free or reduced lunch comprised 26.0% of the sample. The racial composition of the sample was as follows: Caucasian (71.1%), African American (11.7%), More than one race (8.5%), Asian / Pacific Islander (5.9%), and Native American (2.8%). With respect to ethnicity, 23.6% of the children came from Hispanic families. The primary language spoken in the homes of the children included English (64.8%), Spanish (11.9%), and over sixty other languages (23.3%). The children attended kindergarten in all fifty states plus the District of Columbia and Puerto Rico. While this sample is very diverse geographically, it should be noted that 84.6% of the sample came from four states with statewide implementation of the assessment system (Colorado, Delaware, Massachusetts, and Washington).

Results

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 Rasch Rating Scale Model (RSM; Andrich, 1978), with Winsteps software (Linacre, 2012). A separate Rasch analysis was conducted for each of the six domains of development. The RSM and Partial Credit Model (PCM; Masters, 1982) are the two most widely used Rasch model for polytomous response data. The RSM, rather than the PCM, was chosen because the items share the same rating scale structure (i.e., use of the same number of rating scale categories and labels across items). In cases where each item has its own rating scale structure, the PCM would be the appropriate model to apply. This decision was also based on results from nationally representative norm samples, which indicate that the RSM yielded better fit of these data to the model than did the PCM.

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 PCA showed that for the Cognitive scale, the Rasch dimension explained 75.9% of the variance in the data, with an eigenvalue of 31.4. The first contrast (the largest secondary dimension) had an eigenvalue of 1.8 and accounted for only 4.3% of the unexplained variance. The fit statistics for all of the Cognitive items were within acceptable limits: the infit MNSQ ranged from 0.78 to 1.31; the outfit MNSQ ranged from 0.78 to 1.27. The item total score correlations ranged from .79 to .85.

Language Scale (8 items)

The PCA showed that for the Language scale, the Rasch dimension explained 76.6% of the variance in the data, with an eigenvalue of 26.2. The first contrast (the largest secondary dimension) had an eigenvalue of 1.6 and accounted for only for 4.5% of the unexplained variance. The fit statistics for all of the Language items were within acceptable limits: the infit MNSQ ranged from 0.81 to 1.45; the outfit MNSQ ranged from 0.79 to 1.27. The item total score correlations ranged from .76 to .84.

Literacy Scale (12 items)

The PCA showed that the Rasch dimension explained 69.3% of the variance in the data, with an eigenvalue of 27.1. The first contrast (the largest secondary dimension) had an eigenvalue of 2.8 and accounted for 7.2% of the unexplained variance, suggesting some evidence for the possibility of a second underlying construct. The fit statistics for the Literacy items were within acceptable limits: the infit MNSQ ranged from 0.78 to 1.40; the outfit MNSQ ranged from 0.73 to 1.53. The item total score correlations ranged from .61 to .87.

Mathematics Scale (7 items)

The PCA showed that the Rasch dimension explained 71.7% of the variance in the data, with an eigenvalue of 17.7. The first contrast (the largest secondary dimension) had an eigenvalue of 2.0 for 8.1% of the unexplained variance. These results suggest the possibility of a second underlying construct. The fit statistics for the Mathematics items were all within acceptable limits: the infit MNSQ ranged from 0.79 to 1.40; the outfit MNSQ ranged from 0.80 to 1.20. The item total score correlations ranged from .78 to .84.

Physical Scale (5 items)

The PCA showed that for the Physical scale, the Rasch dimension explained 73.9% of the variance in the data, with an eigenvalue of 14.1. The first contrast (the largest secondary dimension) had an eigenvalue of 2.1 and accounted only for 11.2% of the unexplained variance. These results suggest the possibility of a second underlying construct. The fit statistics for all of the Physical items were within acceptable limits: the infit MNSQ ranged from 0.83 to 1.27; the outfit MNSQ ranged from 0.78 to 1.12. The item total score correlations ranged from .76 to .82.

Social Emotional Scale (9 items)

The PCA showed that for the Social Emotional scale, the Rasch dimension explained 75.7% of the variance in the data, with an eigenvalue of 28.1. The first contrast (the largest secondary dimension) had an eigenvalue of 1.8 and accounted only for 4.7% of the unexplained variance. The fit statistics for all of the Social Emotional items were well within acceptable limits: the infit MNSQ ranged from 0.83 to 1.34; the outfit MNSQ ranged from 0.81 to 1.22. The item total score correlations ranged from .76 to .81.

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 rating scale model very well. These results indicated that the data generally satisfied the unidimensionality assumption of the Rasch model. The exceptions to this conclusion where the results suggest the possibility of multiple underlying constructs, or secondary dimensions, for a given scale are tempered by the fact that that all of the item total score correlations are high and all of the fit statistics are within or very close to the acceptable range. These results when taken together indicate that all of the GOLD items share a substantial amount of variance with their respective scale scores.

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 each rating category has a minimum of 10 observations. For all of the items, the teachers used all 10 rating scale points and there were sufficient observations in each of the categories to model the ratings scale. 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 score 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). These results indicate that this was the case for all of the rating scale information across all of the items.

Item Difficulty Measures

For all six scales, the item difficulty or location hierarchy appeared to be generally consistent with the expected developmental trajectory for typically developing kindergarten children. The overall pattern to the item difficulties and the hierarchy of difficulties within each domain is similar to those found in previous studies of the use of GOLD with kindergarten children. The following is a summary for each domain.

Cognitive Scale

The item pertaining to a child's ability to show flexibility and inventiveness in thinking (11.E) was found to be the most difficult item. The item pertaining to a child's ability to engage in sociodramatic play (14.B) was estimated as the easiest item. The range of overall item difficulties (-.43 to .34) is somewhat narrow for ideal separation of children across the range of underlying abilities. However, the range of difficulties is much wider and within the acceptable range when considering the separation created between children by the range of rating scale anchor point locations.

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 engage in conversations (10.A) was estimated as the easiest item. The range of both item difficulties (-.40 to .95) and item anchor point locations was wide enough for reasonable separation of children according to underlying ability.

Literacy Scale

The item pertaining to a child's ability to write to convey meaning (19.B) was found to be the most difficult item. The item pertaining to a child's ability to identify and name letters (16A) was estimated as the easiest item. The range of both item difficulties (-.81 to 1.01) and item anchor point locations was wide enough for reasonable separation of children according to underlying ability.

Mathematics Scale

The item pertaining to a child's ability to compare and measure (22) was found to be the most difficult item. The item pertaining to a child's ability to connect numerals with quantities (20.C) was estimated to be the easiest item. The range of both item difficulties (-.75 to .78) and item anchor point locations was wide enough for reasonable separation of children according to underlying ability.

Physical Scale

The item pertaining to a child's ability to demonstrate gross motor skills (6) was found to be the most difficult item. The items pertaining to a child's traveling skills (4) and ability to use fingers and hands (7.A) were estimated as the easiest items. The range of overall item difficulties (-.24 to .36) is too narrow for ideal separation of children across the range of underlying abilities. However, the range of difficulties is much wider and closer to the acceptable range when considering the separation created between children by the range of rating scale anchor point locations.

Social Emotional Scale

The item pertaining to a child's ability to participate cooperatively in group situations (3.B) 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 (-1.45 to 1.09) and item anchor point locations was wide enough for reasonable separation of children according to underlying ability.

Reliability

Reliability was evaluated using 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 traditional test reliability whereas the item reliability has no traditional 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 1), the scale scores appear to be highly reliable, as evidenced by person separation index of 3.77, person reliability of .93, item separation index of 19.67, and item reliability of .99. The Cronbach's alpha reliability coefficient for this scale was .97, 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 index of 3.34, person reliability of .92, item separation index of 36.69, and item reliability of .99. The Cronbach's alpha reliability coefficient for this scale was .96, 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 index of 3.57, person reliability of .93, item separation index of 66.56, and item reliability of .99. The Cronbach's alpha reliability coefficient for this scale was .96, 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 index of 2.89, person reliability of .89, item separation index of 54.13, and item reliability of .99. The Cronbach's alpha reliability coefficient for this scale was .94, indicating high internal consistency reliability.

Physical Scale

Based on the Rasch reliability indexes, the scale scores appear to be reliable, as evidenced by person separation index of 1.99, person reliability of .80, item separation index of 14.49, and item reliability of .99. The Cronbach's alpha reliability coefficient for this scale was .93, 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 index of 3.09, person reliability of .91, item separation index of 64.10, and item reliability of .99. The Cronbach's alpha reliability coefficient for this scale was .96, indicating high internal consistency reliability.

Scale Scores

The body of evidence to date from research studies and the Rasch modeling using this sample suggests that scale scores for each of the developmental domains outlined by the test developers would be appropriate. 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 raw score to scale conversion table generated by the Rasch modeling software, based on national norm data, was used to rescale these 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.

The scale scores were examined for the degree of correlation between each combination of domain scores. These results are displayed in Table 2. These correlations ranged from .498 to .832. The smallest correlation was between the Physical and Literacy scale scores (.498) and the largest value was between the Cognitive and Language scale scores (.832). These results indicate that while there are moderate to high correlations between the scale scores, most of the correlations fall within the moderate range, supporting the developer's intentions that the six scale scores assess distinct domains of development.

For each scale score, as shown in Table 3, the scale mean, standard deviation, quartile boundaries, and standard error of measurement are reported. The table also indicates the number of children with sufficient data for scale scores were able to be calculated. The standard errors of measurement (SEM) are reported at the scale mean for this sample. In all IRT models, unlike with classical measurement models, the SEM can be estimated for each scale score point. The SEM values may be a little larger for children with scores at the extremely high end of the scale as there is less information available to make estimates for less frequently occurring scores. These results suggest that teachers appear to be discriminating between children across almost the full range of scores. The minimum scores across the scales range from as low as 200 for the Social Emotional and Physical scales to 296 for Literacy. The maximum scale scores are 800 for all six scales. The reported quartile boundaries included can be used to enable teachers to understand approximately where a child's score falls relative to other children in the sample.

Table 4 contains the results of examining the scale scores using the subset of children who were assessed at all three assessment checkpoints (fall, winter, and spring). The mean, standard deviation, and 25th, 50th, 75th percentiles were all calculated for each of the time points as well as for fall to spring growth. These results indicate that GOLD scores are sensitive to the growth that kindergarten children make across the academic year. The growth in average scale scores from fall to spring is around 100 points with a standard deviation around 55 points. Specifically, the children made the following average gains from fall to spring: Cognitive – 96.97 (SD=55.39), Literacy – 102.87 (SD=51.03), Social Emotional – 99.63 (SD=58.01), Physical – 81.88 (SD=62.01), Language – 95.94 (SD=57.08), Mathematics – 102.27 (SD=48.50). Figure 1 illustrates these average growth rates.

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 several nationally representative normative studies. These 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 scales as developmental progressions to place children along a continuum of growth and development. 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 kindergarten classrooms around the country could focus on measures of the degree of association between GOLD scale scores and external measures of child developmental progress. It would also be helpful to conduct inter-rater reliability studies in these settings.

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.

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
Reliability indexes

Domain of developmer	Person Reliability	Person Separation	Item Reliability	Item Separation	Cronbach's Alpha
Cognitive	.93	3.77	.99	19.67	.97
Literacy	.93	3.57	.99	66.56	.96
Social Emotional	.91	3.09	.99	64.10	.96
Physical	.80	1.99	.99	14.49	.93
Language	.92	3.34	.99	36.69	.96
Mathematics	.89	2.89	.99	54.13	.94

Table 2
Correlations between GOLD scale scores by domain of development

Domain of developmen	Literacy	Social Emotional	Physical	Language	Mathematics
Cognitive	0.766	0.816	0.677	0.832	0.771
Literacy		0.650	0.498	0.731	0.830
Social Emotional			0.687	0.744	0.645
Physical				0.590	0.635
Language					0.742

Table 3

Kindergarten entry scores by domain of devleopment

	Cognitive	Literacy	Social Emotional	Physical	Language	Mathematics
Mean	651.27	646.66	630.81	620.61	627.33	641.31
SD	74.51	67.83	69.51	60.37	79.80	64.39
25th	610	607	595	592	580	602
50th	662	651	640	627	639	647
75th	698	694	672	656	682	683
n	49375	76398	57351	63523	65619	30592
SEM	29	31	38	34	39	39

Table 4

Growth across the kindergarten year by domain of development

Domain of development		Fall	Winter	Spring	Growth
Cognitive	Mean	659.88	715.39	756.55	96.67
	SD	73.72	74.18	63.79	55.39
	25th	622.00	679.00	731.00	57.00
	50th	670.00	731.00	785.00	98.00
	75th	706.00	771.00	800.00	130.00
	n	10676	10676	10676	10676
	SEM	29	34	26	
Literacy	Mean	658.19	726.02	761.05	102.87
	SD	67.01	63.73	55.60	51.03
	25th	618.00	694.00	743.00	68.00
	50th	662.00	739.00	782.00	101.00
	75th	705.00	772.00	800.00	135.00
	n	15560	15560	15560	15560
	SEM	33	36	37	
Social Emotional	Mean	637.57	691.49	737.20	99.63
	SD	69.65	66.81	65.58	58.01
	25th	601.00	653.00	696.00	62.00
	50th	647.00	696.00	756.00	101.00
	75th	678.00	732.00	795.00	135.00
	n	13855	13855	13855	13855
	SEM	38	36	27	
Physical	Mean	620.92	664.63	702.80	81.88
	SD	67.11	68.66	69.44	62.01
	25th	592.00	637.00	666.00	49.00
	50th	627.00	666.00	729.00	78.00
	75th	656.00	729.00	729.00	114.00
	n	10003	10003	10003	10003
	SEM	34	34	34	
Language	Mean	635.13	688.03	731.07	95.94
	SD	79.34	78.79	76.94	57.08
	25th	588.00	649.00	697.00	60.00
	50th	649.00	697.00	740.00	95.00
	75th	690.00	740.00	800.00	132.00
	n	13037	13037	13037	13037
	SEM	39	39	26	
Mathematics	Mean	649.99	709.66	752.26	102.27
	SD	61.68	60.21	56.48	48.50
	25th	615.00	676.00	726.00	71.00
	50th	654.00	720.00	764.00	101.00
	75th	691.00	748.00	800.00	132.00
	n	11685	11685	11685	11685
	SEM	39	41	38	

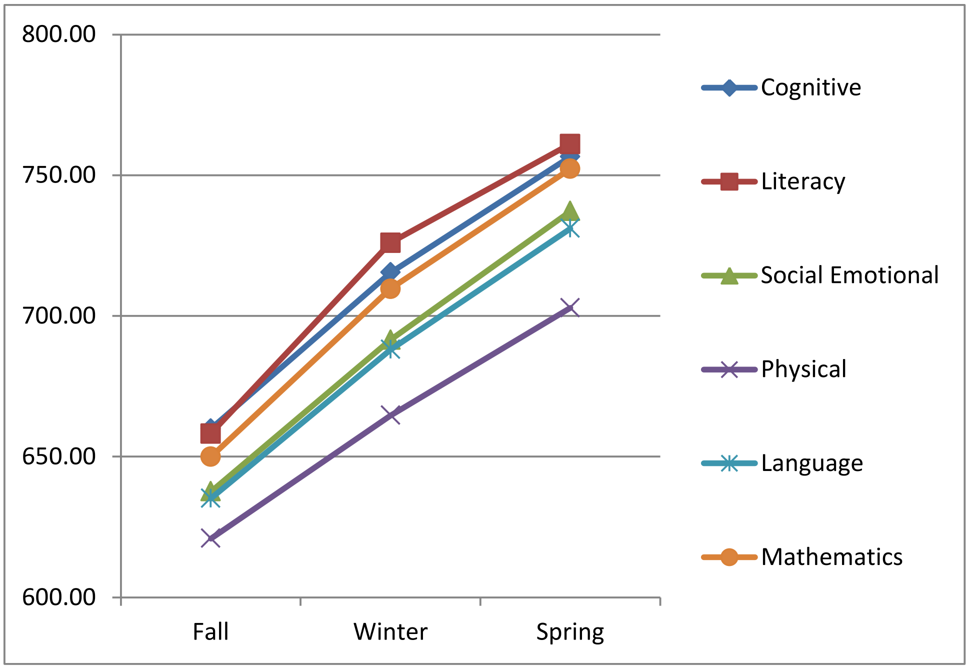

Figure 1. Average fall to spring growth for kindergarten children by scale score.