CEME Technical Report

CEMETR-2010-01
DECEMBER 2010

The Center for Educational Measurement and Evaluation

The Measurement Properties of the *Teaching Strategies GOLD*™ Assessment System: Preliminary Results Following the Winter Assessment Checkpoint April 2010

Richard Lambert
Do-Hong Kim

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

The Measurement Properties of the *Teaching Strategies GOLD™* Assessment System: Preliminary Results Following the Winter Assessment Checkpoint

Richard G. Lambert
Do-Hong Kim
Center for Educational Measurement and Evaluation
UNC Charlotte
April 2010

Sample

The total sample for this second phase of the pilot testing of the *Teaching Strategies GOLD™* assessment system consisted of 2,594 children. The children in this national sample received educational services in 46 different centers that are located in all regions of the United States. All of these centers but one use The *Creative Curriculum®* and had been using *The Creative Curriculum Developmental Continuum for Ages 3-5* assessment system prior to this study. A total of 181 different raters (teachers) provided the ratings for the study. Each teacher received training in the use of the *Teaching Strategies GOLD™* assessment system and rated an average of 14.33 children.

The children span the entire age range for which the assessment system is intended (birth through kindergarten).   It is important to note that the pilot analyses that are discussed in this report were conducted using unweighted data. The sample intentionally included over-sampling of children who are English-language learners.

Validity

Factor analysis. Exploratory factor analysis using maximum likelihood extraction methods, followed by direct oblimin rotations was conducted. A five-factor solution emerged that accounted for approximately 78% of the variance in item responses. This solution very closely matched the organization of the items on the instrument, itself, and maps directly onto the constructs intended by the test development team, as shown below.

a. **Factor 1: Language**
    - All eight language items loaded on Factor 1.

b. **Factor 2: Literacy & Mathematics**
    - Ten of the 12 literacy items (15a, 15b, 16a, 16b, 17a, 17b, 18a, 18b, 18c, 19b) loaded on Factor 2.
        - ✓ Two literacy items (15c and 19a) cross-loaded on Factor 5 (Cognitive) and Factor 3 (Physical), respectively.
    - Six of the seven mathematics items (20a, 20b, 20c, 21b, 22, 23) loaded on Factor 2.
        - ✓ One mathematics item (21a with a weak loading of .35) loaded on Factor 1.

c. **Factor 3: Physical**
    - All 14 physical items loaded on Factor 3.

d. **Factor 4: Social-- Emotional**
    - All nine social-emotional items loaded on Factor 4.

e. **Factor 5: Cognitive**
    - Eight of the 10 cognitive items (11a, 11b, 11c, 11d, 11e, 12b, 13, 14a) loaded on Factor 5.
        - ✓ One cognitive item (14b) has a weak factor loading (<.32)
        - ✓ One cognitive item (12a with a weak loading of .32) loaded on Factor 1.

Rasch Analysis. Data were analyzed using the Rasch Rating Scale Model (RSM; Andrich, 1978), with Winsteps software (Linacre, 2009). The unidimensionality of the scale was evaluated by using Mean Square (MNSQ) item fit statistic 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.

The variance in the data explained by the Rasch measures was 80.9% for Social--Emotional, 82.2% for Physical, 85.5% for Language, 82.9% for Cognitive, 74.2% for Literacy, and 79.1% for mathematics, satisfying the Rasch model for unidimensionality.

The item hierarchy appears to be consistent with the expected developmental trajectory for typically developing children, as shown in the table below. The items are functioning well. For example, no items were found to misfit for the Social--Emotional and Language domains. Only one item was found to misfit for each of the Cognitive, Literacy, and Mathematics domains.

At the second checkpoint, participants were given new items for gross motor Physical development, that expanded teacher's ability to look at specific aspects of the item. These are optional items. The developers recommended that teachers use these items when they had concerns about children's development and needed further clarification and more detailed information than the overall item provided. The developers further recommended that participants use these items with children with delays in physical development. Therefore, the children scored on these items (4A, 4B, 4C, 5A, 5B, 5C, 6A, 6B, 6C) would likely not meet widely held expectations for their age group, which is appropriate. The results of these analyses suggest that these items (4A, 4B, 4C, 5A, 5B, 6A, 6C) appear to be misfit. This issue will be explored further after the final checkpoint.

Item	Item Difficulty	MNSQ Infit	MNSQ Outfit
Social—Emotional			
3A	0.85	0.77	0.76
3B	0.78	0.90	0.90
2C	0.45	1.04	1.04
2B	0.44	0.81	0.78
2D	0.43	1.22	1.22
1A	0.04	1.01	1.03
1B	-0.33	0.87	0.89
1C	-1.04	0.98	0.99
2A	-1.61	1.17	1.15
Physical			
4C	2.86	2.53	2.4
5C	1.65	1.24	1.2
6C	1.4	1.74	1.92
6B	1.33	1.10	1.13
5B	1.13	3.17	3.43
6A	0.49	1.48	1.74
7B	0.13	0.96	0.95
6	-0.01	0.78	0.83
5	-0.46	0.58	0.55
7A	-0.54	0.70	0.73
4	-0.59	0.59	0.62
4B	-0.68	1.75	1.59
5A	-2.76	2.44	2.35
4A	-3.97	1.79	1.54
Language			
9D	1.43	1.06	1.10
9C	0.31	0.88	0.89
10B	0.30	1.04	1.13
10A	-0.06	0.86	0.87
9A	-0.12	0.76	0.78
9B	-0.58	0.76	0.83
8B	-0.62	1.12	1.08
8A	-0.65	0.95	0.95
Cognitive			
13	0.87	1.12	1.08
14A	0.74	0.89	0.86
14B	0.25	1.43	1.47
11E	0.24	0.92	0.89
12A	0.2	0.97	0.97
11C	-0.13	0.87	0.86
11B	-0.36	0.88	0.88
12B	-0.46	0.88	0.85
11D	-0.6	0.85	0.83

11A	-0.76	0.93	0.95
Literacy			
16B	1.15	1.42	1.27
15C	0.64	1.04	1.01
16A	0.13	1.79	1.68
19B	0.1	0.96	0.98
17B	0.09	0.62	0.59
18C	0.07	0.82	0.79
18B	0.05	0.69	0.68
15B	0.05	0.77	0.76
18A	-0.2	0.79	0.78
15A	-0.3	1.00	1.01
19A	-0.76	1.29	1.29
17A	-1.02	0.82	0.93
Mathematics			
20C	0.63	1.49	1.49
22	0.63	0.92	0.96
20B	0.3	0.70	0.70
23	-0.09	1.04	1.03
21A	-0.42	0.96	0.98
20A	-0.44	0.91	0.91
21B	-0.6	0.78	0.76

Note. Misfit in italicized, bold font

The rating scale structure was also examined. Each rating category has the recommended minimum of 10 observations. With a few exceptions, the average measures increase monotonically with the categories. For the Cognitive and Mathematics domains, the average measures do not advance monotonically between category 8 and category 9. The rating categories display adequate fit (i.e., Outfit mean-square values < 2.0), except for category 0 in Physical and category 8 in Cognitive. The thresholds advance with categories (i.e., ordered thresholds) for the domains of Social-- Emotional, Oral Language, and Mathematics. Disordered thresholds are observed for the domains of Physical, Cognitive, and Literacy.

Examination of the person-item map distribution revealed that the items did not cover as wide a range of continuum of the latent variable as might be ideal, suggesting that more items of varying difficulty may be helpful to more closely match the range of developmental levels of the children.

Reliability

Cronbach's alpha reliability coefficients, a measure of the internal consistency of each scale score, was calculated for each of the six domains and found to be high: .966 for Social--Emotional, .968 for Physical, .972 for Language, .976 for Cognitive, .966 for Literacy, and .960 for Mathematics. The Rasch based analyses also produce a number of reliability indexes. Overall, the GOLD assessment system appeared to be highly reliable as indicated by the Rasch reliability statistics, as shown in the table below.

Domain	Person Separation Index	Person Reliability	Item Separation Index	Item Reliability
Social-- Emotional	4.75	0.96	31.56	1.00
Physical	3.36	0.92	22.55	1.00
Language	5.20	0.96	22.60	0.99
Cognitive	5.59	0.97	17.09	1.00
Literacy	4.53	0.95	24.72	1.00
Mathematics	4.11	0.94	18.35	1.00

Summary

Overall, the GOLD assessment system appeared to be highly reliable as indicated by the reliability statistics. Results of the factor analysis show that with very few exceptions, the items loaded onto the constructs intended by the test development team. Our analyses of the dimensionality suggest that the GOLD assessment system measures largely satisfy the Rasch model for unidimensionality.