

Outcomes  
Study

LEVEL OF EVIDENCE  
Gold Standard

*Teaching Strategies GOLD*<sup>®</sup>  
Assessment System

Concurrent Validity

.....  
An Independent Study Comparing External Assessment  
Measures to *Teaching Strategies GOLD*<sup>®</sup>



# Teaching Strategies GOLD® Assessment System

## Concurrent Validity

### **Concurrent Validity of Teaching Strategies GOLD®**

With the virtual explosion of accountability guidelines and standards, child assessment has assumed heightened importance. The National Research Council, the National Association for the Education of Young Children, and the National Association of Early Childhood Specialists in State Departments of Education indicate that assessment measures should be developmentally appropriate, educationally important, and linguistically and culturally responsive. Assessment measures must also demonstrate acceptable psychometric properties of reliability and validity (National Association for the Education of Young Children & National Association of Early Childhood Specialists in State Departments of Education, 2003; Snow & Van Hemel, 2008). Many professionals, including those who work in the field of early intervention, support authentic, observation-based performance assessment of young children (Keilty, LaRocco, & Casell, 2009; McAfee, Leong, & Brodova, 2004). Authentic assessment is ongoing, and information is collected by teachers in typical everyday situations rather than as an add-on to what they are already doing.

---

---

One new authentic assessment measure is *Teaching Strategies GOLD*<sup>®</sup>, and it is the focus of this paper. *Teaching Strategies GOLD*<sup>®</sup> is designed to assess the ongoing development and learning of children from birth through kindergarten, including English-language learners and children with disabilities. Previous studies of *Teaching Strategies GOLD*<sup>®</sup> using large, diverse samples were conducted by the Center for Educational Measurement and Evaluation (CEME) at the University of North Carolina at Charlotte. Researchers found strong evidence for interrater, person and item, and internal consistency reliabilities and for construct validity (Kim, Lambert, & Burts, 2013; Kim & Smith, 2010; Lambert, Kim, & Burts, in press; Lambert, Kim, Taylor, & McGee, 2010). Although these results are indeed encouraging, they do not address the important issue of concurrent validity: how well *Teaching Strategies GOLD*<sup>®</sup> correlates with other measures.

Studies conducted by researchers in Tulsa, Oklahoma, preschool classrooms and by researchers in Washington state kindergarten classrooms explored concurrent validity using *Teaching Strategies GOLD*<sup>®</sup> (Decker, 2013) or a modified version of *Teaching Strategies GOLD*<sup>®</sup> (Soderberg, Stull, Cummings, Nolen, McCutchen, & Joseph, 2013). Results of the Tulsa study (Decker, 2013) indicate that *Teaching Strategies GOLD*<sup>®</sup> demonstrates moderate to high correlations with the Bracken School Readiness Assessment (Panter & Bracken, 2009). Further evidence of concurrent validity was provided in the Washington study with a modified version of *Teaching Strategies GOLD*<sup>®</sup> (i.e., *WaKIDS*). Moderate correlations with a battery of established norm-referenced achievement instruments were found for the language, literacy, and mathematics areas (Soderberg et al., 2013).

### Current Study

The current study further examined the concurrent validity of *Teaching Strategies GOLD*<sup>®</sup>. It explored the associations between *Teaching Strategies GOLD*<sup>®</sup> scale scores and (1) teacher ratings of children's social functioning and learning behaviors and (2) child performance on direct assessments of academic skills.

---

---

## Procedures

The American Institutes for Research (AIR) gathered the data for the study. AIR is an independent, not-for-profit organization that conducts research in the behavioral and social sciences, including large-scale studies of early care and education programs for governmental agencies and private organizations, nationally and internationally (see <http://www.air.org>).

Over a 1-month period, trained AIR data collectors individually administered each of the direct assessment measures. Classroom teachers, who were current *Teaching Strategies GOLD*<sup>®</sup> users, collected the *Teaching Strategies GOLD*<sup>®</sup> assessment information and the data for measures of social functioning and learning behaviors. External assessment instruments were selected to align as closely as possible with domains assessed in *Teaching Strategies GOLD*<sup>®</sup>. All external measures exhibit strong psychometric properties of reliability and validity, are widely used, and have been validated with diverse populations.

## Direct Assessment Measures

The Pre-Language Assessment Scales (Pre-LAS) assesses children's receptive and spoken English abilities. Pre-LAS has scales: Following Directions, Comprehension, Labeling Items, Repeating Sentences, Completing a Sentence, and Retelling a Story. Based on assessment scores, children are assigned to one of three categories: non-English speakers, limited English speakers, or fluent (proficient) English speakers (Duncan & De Avila, 1985).

The Peabody Picture Vocabulary Test, Fourth Edition (PPVT<sup>™</sup>-4), measures receptive vocabulary knowledge in persons from 2-and-a-half years old to adult age (Dunn & Dunn, 2007).

The Woodcock-Johnson<sup>®</sup> III NU Tests of Achievement (W-J III) measures language, literacy, and mathematics skills. Scales used in this study were Oral Expression, Listening Comprehension, Basic Reading Skills, Reading Comprehension, and Mathematics Calculation (Woodcock, McGrew, & Mather, 2007).

The Pencil Tapping task portion of the Preschool Self-Regulation Assessment (PSRA) is a measure of children's inhibitory control (regulation, attention, and behavior). Children are requested to tap their pencil once when the examiner taps twice and to tap it twice when the examiner taps once (Smith-Donald, Raver, Hayes, & Richardson, 2007).

The Head-Toe-Knees-Shoulders task (HTKS) is a measure of preschool and kindergarten children's behavioral regulation whereby children are given various oral instructions (e.g., "touch your toes") and are expected to respond in an atypical way (e.g., children are expected to touch their heads) (Ponitz, McClelland, Matthews, & Morrison, 2009).

---

---

## Teacher Assessments

Children's social skills and learning behaviors are difficult to measure in one testing session by an independent assessor. Therefore, these areas were assessed by children's classroom teachers who were trained to use the PKBS and the PLBS.

The Preschool and Kindergarten Behavior Scales (PKBS, Second Edition) measures the social functioning of children 3 through 6 years old: Social Cooperation, Social Interaction, Social Independence, Externalizing Problems, and Internalizing Problems (Merril, 2003).

The Preschool Learning Behaviors Scale (PLBS) is a 29-item teacher behavior-rating instrument for assessing preschool children's approaches to learning. The measure has three dimensions: Competence Motivation, Attention/Persistence, and Attitudes Toward Learning (McDermott, Leigh, & Perry, 2002).

*Teaching Strategies GOLD*<sup>®</sup> has 36 objectives organized within the areas of social-emotional, physical, language, cognitive, literacy, and mathematics. Teachers gather child information within daily activities as they observe, interact with children and families, and collect examples of children's work (Heroman, Burts, Berke, & Bickart, 2010). Portfolios serve as a basis for rating children along color-coded developmental and learning expectations (see <http://www.teachingstrategies.com>).

## Participants

A stratified random sampling procedure by type of center was used to ensure a sample proportional to the national distribution of Head Start, public school pre-K, and other types of early childhood providers. The sample included 299 children who attended 51 different classrooms across 16 centers located in the northeastern United States. The majority of the classrooms served only 4-year-old children ( $n = 34$ ), some served only 3-year-old children ( $n = 13$ ), and several served both ( $n = 4$ ). There were approximately equal numbers of males and females in the sample. The majority of the children (59%) lived in homes where English was the primary language spoken, though a substantial minority lived in homes where a language other than English was spoken (Spanish: 27%; other: 14%). About one fourth of the children were English-language learners, and about one fourth were from low-income families. Most study children were from minority groups (Hispanic: 45%; African American: 26%). Only children who could be reliably assessed in English were included in the study.

---

---

## Analyses

Pearson correlation coefficients were used to assess the degree of association between the external measures and *Teaching Strategies GOLD*<sup>®</sup> scale scores. Simple Pearson correlation coefficients do not account for the clustering of children within classrooms or the fact that the classroom teachers completed both the *Teaching Strategies GOLD*<sup>®</sup> assessments and the external social skills (PKBS) and learning behaviors (PLBS) assessments. Therefore, a two-level hierarchical linear model (HLM) was created for each correlation coefficient between an external measure and a *Teaching Strategies GOLD*<sup>®</sup> scale score (i.e., Social–Emotional, Cognitive, Language, Literacy, Mathematics). The results of these models were used to estimate the variance in the dependent variable that is accounted for by the respective *Teaching Strategies GOLD*<sup>®</sup> scale score while accounting for teacher rating and clustering effects.

## Results

**Research Question 1:** *What is the degree of association between teacher ratings of the social functioning and the learning behaviors of young children and Teaching Strategies GOLD<sup>®</sup> scale scores?*

*Social Functioning:* It was anticipated that the *Teaching Strategies GOLD*<sup>®</sup> Social–Emotional scale score would be associated with the teacher-rated Social Functioning Assessment scores (PKBS) and would result in the strongest correlations between external measures and *Teaching Strategies GOLD*<sup>®</sup>. This hypothesis was generally confirmed. Moderate correlations were found between *Teaching Strategies GOLD*<sup>®</sup> Social–Emotional scale scores and all the PKBS subscales ( $r = .428$  to  $r = .523$ ) with the exception of the PKBS Internalizing and Externalizing Problem Behaviors which showed weak correlations. The PKBS Social Interaction score was most strongly associated with the *Teaching Strategies GOLD*<sup>®</sup> Cognitive scale score ( $r = .541$ ).

*Learning Behaviors:* It was expected that *Teaching Strategies GOLD*<sup>®</sup> Cognitive scale scores would be associated (and most strongly) with the Learning Behaviors Teacher Rating scores (PLBS). These expectations were generally confirmed. Total PLBS, Persistence, and Motivation were moderately correlated with the *Teaching Strategies GOLD*<sup>®</sup> Cognitive scale scores (range  $r = .428$  to  $r = .486$ ). The Total PLBS score was moderately associated with all *Teaching Strategies GOLD*<sup>®</sup> scale scores ( $r = .426$  to  $r = .507$ ) indicating the importance of learning behaviors to all domains of children’s learning and development.

HLM models indicated that the strength of association was either very similar to or higher than that shown with the Pearson correlation coefficients. The *Teaching Strategies GOLD*<sup>®</sup> Social–Emotional scale score had the strongest association of all *Teaching Strategies GOLD*<sup>®</sup> scale scores for the PKBS and the PLBS, yielding moderately strong associations (17.5% to 41.1% variance accounted for).

---

---

**Research Question 2:** *What is the degree of association between scores on direct assessment measures of the academic skills of young children and Teaching Strategies GOLD® scale scores?*

It was expected that the scale scores from each external measure would be associated with the *Teaching Strategies GOLD®* scale score that measures the most closely related construct and would result in the strongest correlation among those for the *Teaching Strategies GOLD®* scale scores. This hypothesis was generally confirmed.

The PPVT raw score was moderately associated with the *Teaching Strategies GOLD®* Language, Literacy, Mathematics, and Cognitive scale scores ( $r = .436$  to  $r = .483$ ). This is not surprising given the important role vocabulary plays in these academic-related areas. The HTKS score was moderately associated with the *Teaching Strategies GOLD®* Language, Literacy, and Mathematics scale scores ( $r = .356$  to  $r = .389$ ). Both the Pre-LAS and Pencil Tapping scores were moderately associated with all the *Teaching Strategies GOLD®* scale scores (ranges  $r = .307$  to  $r = .412$ ; and  $r = .365$  to  $r = .483$  respectively). For the W-J III measures, Letter-Word Identification, Word Attack, and Understanding Directions scores were moderately associated with the *Teaching Strategies GOLD®* Literacy scale score (range  $r = .372$  to  $r = .450$ ). The Quantitative Concepts score was moderately associated with all the *Teaching Strategies GOLD®* scale scores ( $r = .399$  to  $r = .522$ ) and as expected, the highest value was with the *Teaching Strategies GOLD®* Mathematics scale score.

In the HLM models, for almost every external measure, the *Teaching Strategies GOLD®* Literacy scale score showed the strongest association among the *Teaching Strategies GOLD®* scale scores, yielded a moderately strong association (10.4% to 39.7% variance accounted for), and the strength of association was either very similar to or higher than that shown with the Pearson correlation coefficient. This finding was reasonable given that most of the direct assessment measures either focus on literacy-related constructs or have significant literacy-related components. Some of the associations became notably higher with the HLM models than they were with the simple Pearson correlation coefficients.

---



---

### Summary

Findings from this study support the concurrent validity of *Teaching Strategies GOLD*<sup>®</sup> found in previous studies (Decker, 2013; Soderberg et al., 2013). The correlations of the external measures with the *Teaching Strategies GOLD*<sup>®</sup> domains were for the most part moderate and in expected, aligned areas. Further, this study in general supports findings from studies of other authentic assessments (Meisels, Xue, & Shablott, 2008; Sekino & Fantuzzo, 2005). Moderate level correlations are largely consistent with the expected level of agreement between standardized measures and authentic assessment tools given vast differences in time frame, data sources, and methods for collecting and documenting information between authentic and direct assessments.

The results from this study are encouraging and they add new information concerning the overall psychometric integrity of *Teaching Strategies GOLD*<sup>®</sup> noted in earlier studies (Kim et al., 2013; Kim & Smith, 2010; Lambert et al., in press; Lambert et al., 2010). Taken together, they suggest that programs can avoid duplicative efforts while still assuring stakeholders that *Teaching Strategies GOLD*<sup>®</sup> can be used to gather important information on children's learning and development. Further, the measure is respectful of teachers' and children's time and is more cost-effective for early childhood programs than using multiple measures.

---

---

## References

- Decker, C. G. (2013). *Teaching Strategies GOLD: Testing reliability and validity using the Bracken School Readiness Assessment*. Unpublished report of CAP Tulsa.
- Duncan, S. E., & De Avila, E. A. (1985). *Language assessment scales*. San Rafael, CA: Linguametrics Group.
- Dunn, L. M., & Dunn, D. M. (2007). *Peabody picture vocabulary test* (4th ed.). Bloomington, MN: Pearson.
- Heroman, C., Burts, D. C., Berke, K., & Bickart, T. S. (2010). *Teaching Strategies GOLD® objectives for development and learning*. Washington, DC: Teaching Strategies.
- Keilty, B., LaRocco, D. J., & Casell, F. B. (2009). Early interventionists' reports of authentic assessment methods through focus group research. *Topics in Early Childhood Special Education, 28*(4), 244–256.
- 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*(4), 574–595.
- Kim, D.-H., & Smith, J. D. (2010). Evaluation of two observational assessment systems for children's development and learning. *NHSA Dialog, 13*(4), 253–267.
- Lambert, R. G., Kim, D.-H., & Burts, D. C. (in press). Using teacher ratings to track the growth and development of young children using the Teaching Strategies GOLD® assessment system. *Journal of Psychoeducational Assessment*. doi:10.1177/073482913485214
- Lambert, R. G., Kim, D.-H., Taylor, H., & McGee, J. R. (2010). *Technical manual for the Teaching Strategies GOLD® assessment system*. University of North Carolina at Charlotte: Center for Educational Measurement and Evaluation.
- McAfee, O., Leong, D. J., & Bodrova, E. (2004). *Basics of assessment: A primer for early childhood educators*. Washington, DC: National Association for the Education of Young Children.
- McDermott, P. A., Leigh, N. M., & Perry, M. A. (2002). Development and validation of the preschool learning behaviors scale. *Psychology in the Schools, 39*(4), 353–365.
- Meisels, S. J., Xue, Y., & Shablott, M. (2008). Assessing, language, literacy, and mathematics skills with Work Sampling for Head Start. *Early Education and Development, 19*(6), 963–981.
- Merrell, K. W. (1994). *Preschool and kindergarten behavior scales: Test manual*. Brandon, VT: Clinical Psychology Publishing Company. Abstract retrieved from <http://eric.ed.gov/?id=ED378210>
- National Association for the Education of Young Children & National Association of Early Childhood Specialists in State Departments of Education. (2003). *Early childhood curriculum, assessment, and program evaluation: Building an effective, accountable system in programs for children birth through age 8* [Joint position statement]. Retrieved from <http://www.naeyc.org/files/naeyc/file/positions/CAPEexpand.pdf>
- Panter, J. E., & Bracken, B. A. (2009). Validity of the Bracken School Readiness Assessment for predicting first grade readiness. *Psychology in the Schools, 46*(5), 397–409.
- Ponitz, C. C., McClelland, M. M., Matthews, J. S., & Morrison, F. J. (2009). A structured observation of behavioral self-regulation and its contribution to kindergarten outcomes. *Developmental Psychology, 45*(3), 605–619.
-

Sekino, Y., & Fantuzzo, J. (2005). Validity of the Child Observation Record: An investigation of the relationship between COR dimensions and social-emotional and cognitive outcomes for Head Start children. *Journal of Psychoeducational Assessment, 23*(3), 242–260.

Smith-Donald, R., Raver, C. C., Hayes, T., & Richardson, B. (2007). Preliminary construct and concurrent validity of the Preschool Self-regulation Assessment (PSRA) for field-based research. *Early Childhood Research Quarterly, 22*(2), 173–187.

Snow, C. E., & Van Hemel, S. B. (Eds.). (2008). *Early childhood assessment: Why, what, and how*. Washington, DC: National Academies Press.

Soderberg, J. S., Stull, S., Cummings, K., Nolen, E., McCutchen, D., & Joseph, G. (2013). *Inter-rater reliability and concurrent validity study of the Washington Kindergarten Inventory of Developing Skills (WaKIDS)*. Unpublished report prepared for the State of Washington Office of Superintendent of Public Instruction.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2007). *Woodcock-Johnson® III normative update*. Rolling Meadows, IL: Riverside.

---

