CEMETR-2013-05
NOVEMBER 2013

CEME Technical Report

The Center for Educational Measurement and Evaluation

Evidence for the Association between Scores from the Teaching Strategies GOLD® Assessment System and Information from Direct Assessments of Child Progress

Richard Lambert
Do-Hong Kim
Diane C. Burts

A PUBLICATION OF THE CENTER FOR EDUCATIONAL MEASUREMENT AND EVALUATION

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

Abstract

This study examined the associations between scale scores obtained from a teacher observation-based authentic assessment measure, the *Teaching Strategies GOLD®*, and (a) teacher ratings of children's social functioning and their learning behaviors and (b) child performance on individually administered direct assessments of academic skills. The sample was diverse and included 299 preschool children attending 51 different Head Start, public pre-k, and private school classrooms across 16 centers in the Northeast United States. Pearson correlation coefficients and a two level Hierarchical linear model (HLM) were used to assess the degree of association between *GOLD®* scale scores and external measures. The correlations of the external measures with the *GOLD®* domains were generally moderate and in expected, aligned areas. Findings suggest that the *GOLD®* is a viable authentic assessment measure that taps preschool

Concurrent Validity of the *Teaching Strategies GOLD*®

Introduction

With the virtual explosion of accountability guidelines and standards, child assessment has assumed heightened importance. Well-designed, scientifically-informed assessment measures are necessary to ensure that all children, regardless of culture, language, or disabilities are assessed accurately and fairly (Hirsh-Pasek, Kochanoff, Newcombe, & de Villiers, 2005; Snow & Van Hemel, 2008). 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 National Association of Early Childhood indicate that assessment measures should be developmentally appropriate, educationally important, and linguistically and culturally responsive. These groups, along with the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) emphasize that measures must be used only for the purpose/s for which they were designed (AERA, APA, & NCME, 1999). Moreover they note that assessment instruments must demonstrate acceptable psychometric properties of reliability and validity (AERA, et al., 1999; NAEYC & NAECS/SDE, 2003 Snow & Van Hemel, 2008). The purpose of this paper is to further explore the validity of a relatively new authentic assessment tool, the *Teaching Strategies GOLD*® (*GOLD*®) (Heroman, Burts, Berke, & Bickart, 2010). It examined the associations between *GOLD*® scale scores and (a) teacher ratings of children's social functioning and their learning behaviors and (b) child performance on direct assessments of academic skills.

Authentic Assessment

Many professionals support authentic, observation-based performance assessment of young children (e.g., Copple & Bredekamp, 2009; Gullo, 2006; Keilty, LaRocco, & Casell, 2009; McAfee, Leong, & Brodova, 2004). Authentic assessment provides information that is useful to teachers for instructional planning, individualizing instruction, and communicating with families and other stakeholders. In authentic assessment, evaluation is ongoing, and information is collected by teachers during typical everyday situations rather than as an add-on to what they are already doing. Teachers observe and document what children say and do, select suitable examples and artifacts that illustrate particular abilities and knowledge, and interpret and use the information as they plan and communicate child progress. Capturing children's emerging abilities and their performance as they engage in the active process of learning provides insights that may not be gleaned from single mode or one assessment setting.

Although there is much support for authentic assessment, the research basis for this type of assessment is fairly limited (Hallam, Grisham-Brown, Gao, & Brookshire, 2007). One question often asked by stakeholders concerning authentic assessment measures relates to concurrent validity, that is, validity evidence based on relations to other measures (AERA, et al., 1999; Messick, 1995). They want to know how the information gathered using the authentic assessment measure compares with child data obtained via accepted, traditional, direct assessments.

Several performance-based, authentic assessment instruments are currently used in early childhood classrooms. The measures differ in several respects from one another and from the *GOLD*® including the domains measured, the age ranges for which they are intended, their psychometric properties, and the samples used in validation studies. The

HighScope *Child Observation Record* (*COR*) is designed to measure the development of children ages 2 ½ years to 6 (Schweinhart, McNair, Barnes, & Larner, 1993); Infant-Toddler *COR* is used to assess children ages 6 weeks to 3 years (High/Scope Educational Research Foundation, 2012). The *Learning Through Relating* (*L-TR-CAR*) is designed for use with infants and toddlers (Moreno & Klute, 2011). Meisels and colleagues developed several authentic measures. The *Work Sampling System* (*WSS*) is a curriculum-embedded performance assessment of children preschool to fifth grade (Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001; Meisels, Jablon, Marsden, Dichtelmiller, & Dorfman, 2001; Meisels, Liaw, Dorfman, & Nelson, 1995). The *Work Sampling for Head Start* (*WSHS*) is designed to be used with 3-and 4-year-old Head Start children (Miesels, Xue, & Shamblott, 2008), and the *Ounce Scale* is intended to be used with infants and toddlers from birth to 42 months of age (Meisels, Wen, & Beachy-Quick, 2010. Overall, moderate to strong psychometric properties are reported for the assessment tools. Validation studies of the measures were generally conducted using relatively small samples from Early Head Start, Head Start, or urban school districts located in limited geographic areas (see individual studies for detailed information on each instrument).

One new authentic assessment measure is the *Teaching Strategies GOLD®* (*GOLD®*). The *GOLD®* is intended to assess the ongoing development and learning of children birth through kindergarten and to be inclusive of English language learners (ELLs) and children with disabilities. Previous studies of the *GOLD®* using large, diverse samples were conducted by the Center for Educational Measurement and Evaluation (CEME) - The University of North Carolina at Charlotte. A total of 111,059 children (birth – 71 months) were rated by their classroom teachers using the *GOLD®* during the fall of 2010. From

these, a norm sample of 10,963 children was created corresponding to the 2009 U.S. Census Bureau estimates for children birth to 71 months with regard to ethnicity, gender, geographical region, and state. A similar procedure was used to create a second, longitudinal norm sample (n=20,970) from among all children who were rated at all three checkpoints (i.e., fall, winter, spring) during the 2010-2011 academic year. Thorough descriptions of the sampling procedures are reported in separate manuscripts (names omitted for blind review). Researchers found strong evidence for interrater, person and item, and internal consistency reliabilities and for construct validity (names omitted for blind review). Although these results are hold promise for the measure as a reliable and valid authentic assessment, they do not address the important issue of concurrent validity - how well the *GOLD*® correlates with other measures. This information is especially critical given the widespread use of the *GOLD*® in Head Start and other early childhood programs (J. Mosley, personal communication, September 26, 2013).

Studies conducted by researchers in Tulsa with preschool children (n=1,100) and by researchers in Washington state with kindergarten children (n=333) explored the concurrent validity of the *GOLD*® (Decker, 2013) or a modified version of the *GOLD*® (Soderberg, Stull, Cummings, Nolen, McCutchen, & Joseph, 2013). Results of the Tulsa study (Decker, 2013) indicated that the *GOLD*® demonstrates moderate to high correlations in Language (r =.64), Cognitive (r = .64) Mathematics (r =.74) areas with raw scores on the *Bracken School Readiness Assessment* (Panter & Bracken, 2009). Further evidence of concurrent validity was provided in the Washington study using a modified version of the *GOLD*® (i.e., *WaKIDS*). Moderate correlations (r =.50 - .64) 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 the *Teaching Strategies GOLD®*. More specifically, the study addressed the following research questions: (1) What is the degree of association between teacher ratings of the social functioning and the learning behaviors (i.e., approaches to learning) of young children and GOLD® scale scores? and (2) What is the degree of association between scores on direct assessment measures of the academic skills of young children and GOLD® scale scores?

Method

Procedures

The American Institutes for Research (AIR) collected the data for the study. AIR is an independent, not-for-profit organization that undertakes 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 www.air.org).

Over a one-month period, trained AIR data collectors individually administered each of the external direct assessment measures. Children were taken to a quiet room separate from their classrooms where they were assessed during 35–45 minute sessions twice during the four-week testing window. No individual assessment lasted the full length of time, but the allotted period allowed for warm-up, multiple assessments, and short breaks. Classroom teachers, who were current *GOLD®* users, collected the *GOLD®* assessment information and the data from measures of social functioning and learning behaviors. External assessment instruments

were selected by AIR to align as closely as possible with domains assessed in the *GOLD*®. 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. The measure has six scales: Simon Says for following directions, Choose a Picture to assess comprehension, What's in the House to label items, Say What You Hear to repeat sentences, Finish Stories to complete a sentence, and Let's Tell Stories to retell a story that was just told. Based on raw assessment scores, children are assigned to one of three categories; Non-English Speakers, Limited English Speakers, and Fluent (Proficient) English Speakers (Duncan, & DeAvila, 1985). The *PreLAS* was normed on 850 examinees at nine different sites, and demonstrates strong internal consistency reliability (Cronbach's Alpha is .80 –.90 for the subtests and the part/whole correlations). Only children who were assigned as Limited English Speakers or Fluent/Proficient English Speakers were included in the present study. These criteria provided diversity in the sample while ensuring that children could be reliably assessed with the other measures described as follows.

The *Peabody Picture Vocabulary Test, Fourth Edition* (*PPVT*®-*4*) measures receptive vocabulary knowledge in persons 2.5 years to adults (Dunn & Dunn, 2007). The measure was normed on approximately 3,500 subjects who matched the U.S. census for gender, race/ethnicity, region, socioeconomic status (SES), and clinical diagnosis or special education placement. Validity and reliability coefficients are in the .90 range.

The *Woodcock-Johnson III NU Tests of Achievement* (*W-J III*) measures language, literacy, and mathematics skills (Woodcock, McGrew, & Mather, 2007. Scales used in this study were Oral Expression, Listening Comprehension, Basic Reading Skills, Reading Comprehension, and Mathematics Calculation (Woodcock et al.,, 2007). The WJ-III was normed on a large, representative sample and has high reliability (r = .90 or higher). Scales include Oral Expression, Listening Comprehension, Basic Reading Skills, Reading Comprehension, Mathematics Calculation, and Written Expression. The written expression portions of the measure were not included in this study.

The *Pencil Tapping* task portion of the *Preschool Self-Regulation Assessment-PSRA* (Smith-Donald, Raver, Hayes, & Richardson, 2007) is a measure of children's inhibitory control (regulation, attention, and behavior). It requires children to remember the rule for the correct-response while inhibiting their natural inclination to imitate the experimenter's actions. For example, the child is told to tap the pencil once when the examiner taps twice and to tap it twice when the examiner taps once. The task has been validated (Smith-Donald et al., 2007) and used in various studies of young children's self-regulation (Blair & Razza, 2007; Brock, Rimm-Kaufman, Nathanson, & Grimm, 2009).

The *Head-Toe-Knees-Shoulders Task* (*HTKS*) (Ponitz, McClelland, Matthews, & Morrison, 2009) is a measure of preschool and early elementary children's behavioral regulation. The measure is a more complex, extended version of the *Head-to-Toes* (*HTT*) task (Ponitz, McClelland, Jewkes, Connor, Farris, & Morrison, 2008) whereby children are given various paired oral behavioral instructions (e.g., "touch your toes") and are expected to respond in an atypical way (e.g., child touches head). The measure demonstrated adequate validity, reliability, and variability in scores (Ponitz et al., 2009).

Teacher Assessment Measures

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 instruments. Teachers had previously been trained to use the *Teaching Strategies GOLD*® and were current users of the assessment system.

External measures. The *Preschool and Kindergarten Behaviors Scales (PKBS- second edition*) measures the social functioning of children three through six years: Social Cooperation, Social Interaction, Social Independence, Externalizing Problems and Internalizing Problems (Merril, 2003). The PKBS was normed on a sample of 3,317 children ages 3 to 6 years that is similar to the general U.S. population in the 2000 census in ethnicity, socioeconomic status, and special education classification. The internal consistency reliability for PBKS ranges from .96 to .97 for the two scale totals, and .81 to .95 for the subscales.

The *Preschool Learning Behaviors Scale* (*PLBS*) is a 29-item teacher behavior rating instrument for assessing preschool children's approaches to learning (McDermott, Leigh, & Perry, 2002). The measure was normed on a national sample of 100 3- to 5 1/2-year-olds (McDermott et al., 2002). The *PLBS* was found to have three distinct and reliable dimensions (Fantuzzo, Perry, & McDermott, 2004; McDermott, et al., 2002): Competence Motivation ($\alpha = .85$), Attention/Persistence ($\alpha = .83$), and Attitude Toward Learning ($\alpha = .75$). Dimensions were consistent across child sex, age, ethnicity, and parent education level and in a Head Start sample. Concurrent validity points to the *PLBS* being positively related

to positive social skills, negatively related to behavior problems, and not related to cognitive ability (Fantuzzo et al., 2004).

The Teaching Strategies GOLD® (*GOLD®*). The *GOLD®* (Heroman et al., 2010) is intended to assess the development and learning of children birth through kindergarten and to assist teachers in planning and individualizing instruction and in monitoring and communicating child progress with families and other stakeholders. Although it is closely linked to *The Creative Curriculum® for Infants, Toddlers, & Twos* and *The Creative Curriculum® System for Preschool* (Teaching Strategies LLC, n.d.), its creators intend that the measure can be used in programs that do not use the curricula as well as those programs that do use the curricula developed by Teaching Strategies. The *GOLD®* is widely used in all states for Pre-k assessment. Additionally, Teaching Strategies, LLC has 22 state-level agreements for Pre-k assessment and 12 state-level agreements for kindergarten assessments (J. Mosley, personal communication, September 26, 2013).

Development of the *GOLD®* occurred over several years and incorporated comments from teachers, administrators, consultants, and Teaching Strategies, LLC professional-development staff. Two pilot studies were conducted with diverse populations, and national experts (e.g., development, content areas, special needs, ELLs) provided content review. Final assessment items resulted from the feedback received during the development process; consideration of state early learning standards and the *Head Start Child Development and Early Learning Framework* (U.S. Department of Health & Human Services, 2010); and professional literature and current research identifying the knowledge, skills, and behaviors most predictive of school success.

The *GOLD*® has 36 research-based objectives organized within the areas of Social-Emotional, Physical, Language, Cognitive, Literacy, and Mathematics. Two additional objectives relate specifically to English language acquisition (receptive and expressive) but were not included in this study. The *GOLD*® objectives help teachers focus the assessment process as they regularly gather information through observations, conversations, artifacts, etc. during daily activities. Documentation may be gathered using various means such as audio or video recordings, photographs, or observational notes and then later entered into the assessment system. Child assessment information is summarized at three checkpoint periods (i.e., fall, winter, and spring) using paper or online versions of the instrument.

Teachers use the accumulated information to rate each child's skills, knowledge, and behaviors along a 10- point progression of development and learning from "Not Yet" (Level 0) to Level 9 (child exceeds kindergarten-level expectations). Levels 2, 4, 6, and 8 are "Indicators" and include examples of observable behaviors which help teachers assess child progress toward the objective. Additional steps in the progression, the "In-Between Levels," capture the nuances in children's development and learning and denote that the child's skills in the area are emerging but aren't established. These levels (i.e., 1, 3, 5, 7, and 9) do not include examples and indicate that teacher support (e.g., physical, visual, verbal, gestures, modeling) may be needed to help the child accomplish the objective. Overlapping, color-coded bands indicate the typical age and/or grade-level (i.e., kindergarten) ranges for each item measured.

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 (e.g., private centers). The sample included 299 preschool children who attended 51 different classrooms across 16 centers located in three states in the Northeast region of the United States. The majority of the classrooms served only four-year-old children (n=34), some served only three-year-olds (n=13), and several served both (n=4). There was 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%). Table 1 contains information regarding the demographics of the children in the sample.

Analyses

Pearson correlation coefficients were used to assess the degree of association between external measures and *GOLD*®scale scores (e.g., Social-Emotional, Cognitive, and Literacy). It is important to note that the simple Pearson correlation coefficients do not account for the clustering of children within classrooms. The children are nested, or clustered, within classrooms and the classroom teacher completed both the *GOLD*® assessments and the social skills and learning behaviors rating scales. Therefore, a two level Hierarchical linear model (HLM) was created for each association between scores from external measures and *GOLD*® scale scores to account for the nesting effects of one rater for each classroom of children. For the HLM models, the sample was restricted to a subset of classrooms that contributed at least five children to the sample in order to ensure adequate estimation of within classroom variance. This criterion reduced the sample to

251 children in 33 classrooms. The majority of this subset of classrooms served only four year old children (n=23), some of these classrooms served only three year olds (n=7), and several served both (n=3). The HLM models took the following form:

Level-1 Model

$$Score_{ij} = \beta_{0j} + \beta_{1j}*(GOLD\ scale\ score_{ij}) + r_{ij}$$

Level-2 Model

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

The results of these models were used to examine the relationship between external measures and the *GOLD*® scale scores while accounting for the teacher rating and clustering effects. Each level one model contained the score for one of the external measures as the dependent variable and a score for one of the *GOLD*® scale scores as the single explanatory variable. The results of these models were used to estimate the variance in the dependent variable that is accounted for by the respective *GOLD*® scale score. This variance accounted for statistic was calculated in several steps. First, a null or unconditional model that did not contain the explanatory variable was created for each dependent variable. The variance of the dependent variable was decomposed into the portion that was due to between classroom variance and the portion that was due to within classroom variance.

The Intraclass Correlation Coefficient (ICC) was also calculated using these same variance components and in this application was equal to the proportion of the total variance in the dependent variance that was comprised of between classroom variance. The ICC values for each dependent variable are reported in Tables 3 and 5. Next, the explanatory variable, in this case the *GOLD*® scale score in question, was added to the model. In general, if an explanatory variable is related to the dependent variable, then the within classroom residual variance in the conditional model should decrease. The estimated variance accounted for statistic was therefore calculated by subtracting the remaining within classroom residual variance in the dependent variable, after the addition of the explanatory variable, from the total within classroom residual variance from the null model, and then dividing by the total within classroom variance from the null model. These values are reported in Tables 3 and 5 and can be used to assess the degree of association between scores from external measures and *GOLD*® scale scores given that the nesting within classrooms and teachers has been accounted for.

The results section that follows is organized according to the research questions and to allow the reader to consider the statistical evidence from both the simple Pearson correlation coefficients (Tables 2 and 4) and the model-based estimates of association from the HLM models (Tables 3 and 5). The model-based estimates can be interpreted in a similar fashion to r2 values and therefore represent the proportion of the variance in each external measure that can be accounted for by each of the *GOLD*® scale scores. Similarly, the square root of these values can be interpreted in a similar fashion to r values. It is important to note that two methods will not agree exactly. The HLM models account for the fact that the children are naturally nested in clusters, or classrooms, and are rated by a

single rater within their respective classrooms. There is between classroom variability that may be related to rater effects and or between classroom differences in child demographics and ability levels on the constructs being assessed. The ICC values included on Tables 3 and 5 quantify how much between classroom variance there is for each external measure.

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 GOLD® scale scores?*

Social Functioning: It was anticipated that the *GOLD*® 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 *GOLD*®. This hypothesis was generally confirmed. Moderate correlations were found between *GOLD*® 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 *GOLD*® Cognitive scale score (r =.541) (see Table 2). Values greater than .400 are bolded on Tables 2 and 4 to highlight correlation coefficients that are at least moderate in strength.

Learning Behaviors: It was expected that *GOLD*® Cognitive scale scores would be associated (and most strongly) with the learning behaviors teacher rating scores (*PLBS*). These expectations were generally confirmed as presented in Table 2. Total *PLBS*, Persistence, and Motivation were moderately correlated with the *GOLD*® Cognitive scale scores (range r = .428 to r = .486). The Total *PLBS* score was moderately associated with all *GOLD*® 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 (see Table 3). Values greater than .160 are bolded on Tables 3 and 5 to highlight variance accounted for statistics that could be interpreted as approximately equivalent to correlation coefficients of .400 (i.e., of at least moderate in strength). The *GOLD*® Social Emotional scale score had the strongest association of all *GOLD*® 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 GOLD® scale scores?*

It was expected that the scale scores from each external measure would be associated with the *GOLD*® scale score that measures the most closely related construct and would result in the strongest correlation among those for the *GOLD*® scale scores. This hypothesis was generally confirmed as presented in Table 4.

The PPVT raw score was moderately associated with the *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 *GOLD*® Language, Literacy, and Mathematics scale scores (r = .356 to r =.389). Both the ***Pre-LAS*** and ***Pencil Tapping*** were moderately associated with all of the *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 *GOLD*® Literacy scale score (range r = .372 to r = .450) . The Quantitative Concepts score was moderately associated with all the *GOLD*® scale scores (r = .399 to r =.522) and as expected, the highest

value was with the *GOLD*® Mathematics scale score. The Applied Problems score had its highest association with the *GOLD*® Mathematics scale score, but the association was weak (r = .287).

In the HLM models, for almost every external measure, the *GOLD*® Literacy scale score showed the strongest association among the *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 (see Table 5).

Summary and Discussion

Findings from this study support the concurrent validity of the *Teaching Strategies GOLD*® found in previous studies (Decker, 2013; Soderberg et al., 2013). The correlations of the external measures with the *GOLD*® domains were for the most part moderate and in anticipated, aligned areas. 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 (e.g., Meisels, Xue, & Shamblott, 2008; Sekino & Fantuzzo, 2005; Soderberg, et al., 2013).

The results from this study add new information concerning the overall psychometric integrity of the *GOLD*® (names omitted for blind review). Within the

limitations of this study, the findings imply that the *GOLD*® successfully taps areas assessed using established external measures. Taken together with previous research, findings suggest that the *GOLD*® is a valid and reliable authentic assessment tool. As an integrated part of daily activities, the measure is unobtrusive and respectful of teachers' and children's time. Further, because the *GOLD*® taps areas assessed by other measures, programs can avoid costly, duplicative efforts while still assuring stakeholders that the measure can be used to gather important information on children's learning and development.

Although the results from this study are promising, research on the psychometric properties of the *GOLD*® should continue as long as the instrument is in use. This is crucial given its high visibility and reported widespread use. The early childhood community would benefit from studies which further explored the concurrent validity of the *GOLD*®. Additional research should compare children's development and learning as measured by the *GOLD*® with scores obtained from other assessment instruments in addition to the ones included in the present study. Although the sample was diverse, representative samples should be drawn from other geographic areas of the United States and include children of all the ages assessed with the measure. This would add further credibility to the use of the *GOLD*® with varied populations and in dissimilar geographic regions.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999) *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Blair, C., & Razza, R. P., (2007). Relating effortful control, executive function, and false-belief understanding to emerging math and literacy ability in kindergarten. *Child Development, 78,* 647-663.

Brock, L., Rimm-Kaufman, S. E., Nathanson, L., & Grimm, K. J. (2009). The contribution of 'hot' and 'cool' executive function to children's academic achievement, learning-related behaviors, and engagement in kindergarten. *Early Childhood Research Quarterly, 24*, 337-349.

Copple, C., & Bredekamp, S. (Eds.). (2009). *Developmentally appropriate practice in early childhood programs serving children from birth to age 8* (3rd ed.). Washington, DC: National Association for the Education of Young Children.

Decker, C. G. (2013). *Teaching Strategies GOLD: Testing reliability and validity using the Bracken School Readiness Assessment*. Unpublished report of the CAP, Tulsa, OK.

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.

Fantuzzo, J., Perry, M. A., & McDermott, P. (2004). Preschool approaches to learning and theirrelationship to other relevant classroom competencies for low-income children. *School Psychology Quarterly, 19*(3), 212-230.

Gullo, D. F. (2006). Assessment in kindergarten. In D.F. Gullo (Ed.), *K today: Teaching and learning in the kindergarten year* (pp. 138 – 147). Washington, DC: National Association for the Education of Young Children.

Hallam, R., Grisham-Brown, J., Gao, X. & Brookshire, R. (2007). The effects of outcomes-driven authentic assessment on classroom quality. *Early Childhood Research and Practice, 9*(2). Retrieved December 27, 2011, from http://ecrp.uiuc.edu/v9n2/hallam.html

Heroman, C., Burts, D. C., Berke, K., & Bickart, T. S. (2010). *Teaching Strategies GOLD® objectives for development and learning*. Washington, DC: Teaching Strategies, LLC.

Hirsh-Pasek, K., Kochanoff, A., Newcombe, N. S., & de Villiers, J. (2005). Using scientific knowledge to inform preschool assessment: Making the case for "empirical validity." *Social Policy Report, 14*(1), 1-19.

High/Scope Educational Research Foundation (2012). *Infant-Toddler COR Appendix B: Development and validation*. Retrieved February 6, 2012 from www.highscope.org/Content.asp?ContentId=85

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-K., 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.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

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., Bickel, D. D., Nicholson, J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment in kindergarten to grade 3. *American Educational Research Journal, 38*(1), 73-95.

Meisels, S. J., Jablon, J. R., Marsden, D. B., Dichtelmiller, M. K., & Dorfman, A. B. (2001). *The work sampling system*. San Antonio, TX: Pearson.

Meisels, S. J., Liaw, F., Dorfman, A., & Nelson, R. F. (1995). The work sampling system: Reliability and validity of a performance assessment for young children. *Early Childhood Research Quarterly, 10,* 277-296.

Meisels, S. J., Wen, X., & Beachy-Quick, K. (2010). Authentic assessment for infants and toddlers: Exploring the reliability and validity of the Ounce Scale. *Applied Developmental Science, 14,* 55-71.

Meisels, S. J., Xue, Y., & Shamblott, M. (2008). Assessing, language, literacy, and mathematics skills with *Work Sampling for Head Start*. *Early Education and Development, 19*(6),

963-981.

Moreno, A. J., & Klute, M. M. (2011). Infant-toddler teachers can successfully employ authentic assessment: The *Learning Through Relating* system. *Early Childhood Research Quarterly, 26*, 484-496.

NAEYC & NAECS/SDE (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 January 4, 2013, from www.naeyc.org/cape

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.

Schweinhart, L. J., McNair, S., Barnes, H., & Larner, M. (1993). Observing young children in action to assess their development: The High/Scope Child Observation Record Study. *Educational and Psychological Measurement, 53*, 445-455.

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.

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

Soderberg, J., 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 Washington State Office of Superintendent of Public Instruction.

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*, 173-187.

U.S. Department of Health & Human Services, Administration for Children and Families, Office of Head Start (2010). *The Head Start child development and learning framework: Promoting positive outcomes in early childhood programs serving children 3-5 yearsold*. Washington, DC: Author.

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

Table 1.
Demographic composition of the sample of participants.

Variable	Categories	Percent
Gender	Female	53.3%
	Male	46.7%
Language spoken in the home	English	58.9%
	Spanish	27.1%
	Other	14.0%
English language learner	Yes	24.9%
	No	75.1%
Eligible for free or reduced lunch	Yes	23.5%
	No	76.5%
Ethnicity	White	9.0%
	African-American	26.4%
	Native American	0.3%
	Asian	7.4%
	Multi-racial	12.0%
	Hispanic	44.8%

Note. n=299.

Table 2.
Correlation between teacher ratings and GOLD scale scores.

Measure	GOLD Scale Score				
	Cognitive	Literacy	Social Emotional	Language	Mathematics
PKBS Social Cooperation	**0.517**	0.383	**0.523**	**0.400**	**0.401**
PKBS Social Interaction	**0.474**	0.328	**0.428**	0.353	0.339
PKBS Social Independence	**0.541**	0.392	**0.517**	**0.418**	**0.403**
PKBS Problem Behaviors - Externalizing Problems	-0.280	-0.118	-0.346	-0.185	-0.175
PKBS Problem Behaviors - Internalizing Problems	-0.255	-0.091	-0.291	-0.168	-0.163
PLBS Motivation	**0.428**	0.384	**0.419**	0.394	0.387
PLBS Persistence	**0.483**	**0.441**	**0.494**	**0.425**	**0.439**
PLBS Attitude	0.395	0.286	**0.464**	0.321	0.289
PLBS Total	**0.486**	**0.427**	**0.507**	**0.433**	**0.426**

Note. Values greater than or equal to $r = .400$ are bolded. All r values greater .189 are statistically significant at $p < .001$ for $n = 299$.

Table 3.
Model-estimated variance accounted for by GOLD scale scores for each teacher rating measure.

Measure	ICC	GOLD Scale Score				
		Cognitive	Literacy	Social Emotional	Language	Mathematics
PKBS Social Cooperation	0.192	**0.200**	0.144	**0.248**	0.147	0.120
PKBS Social Interaction	0.352	**0.258**	**0.189**	**0.253**	**0.211**	**0.190**
PKBS Social Independence	0.373	**0.240**	0.159	**0.253**	**0.187**	0.141
PKBS Problem Behaviors - Externalizing Problems	0.292	0.083	0.067	**0.175**	0.036	0.021
PKBS Problem Behaviors - Internalizing Problems	0.281	**0.307**	**0.239**	**0.307**	**0.258**	**0.294**
PLBS Motivation	0.140	**0.329**	**0.279**	**0.333**	**0.235**	**0.240**
PLBS Persistence	0.164	**0.329**	**0.302**	**0.347**	**0.197**	**0.222**
PLBS Attitude	0.154	**0.162**	0.105	**0.259**	0.126	0.144
PLBS Total	0.182	**0.359**	**0.306**	**0.411**	**0.248**	**0.253**

Note. Values greater than or equal to .160 are bolded and equate to an approximate *r* value of .400 or greater.

Table 4.
Correlation between direct assessment measures and GOLD scale scores.

Measure	GOLD Scale Score				
	Cognitive	Literacy	Social Emotional	Language	Mathematics
Peabody Picture Vocabulary Test - Raw Score	**0.436**	**0.456**	0.318	**0.478**	**0.483**
Peabody Picture Vocabulary Test - Standard Score	0.135	0.182	0.034	0.198	0.212
Head-Toes-Knees-Shoulders (HTKS)	0.330	0.356	0.283	0.389	0.365
Pencil Tapping	**0.402**	**0.429**	0.365	**0.441**	**0.483**
Pre-LAS	**0.404**	**0.412**	0.307	0.378	**0.404**
WJ Letter-Word Identification Age Equivalent Score	0.310	**0.450**	0.228	0.274	**0.456**
WJ Understanding Directions Age Equivalent Score	0.311	0.372	0.243	0.341	0.341
WJ Applied Problems Age Equivalent Score	0.221	0.270	0.135	0.261	0.287
WJ Word Attack Age Equivalent Score	0.312	0.379	0.267	0.289	**0.420**
WJ Quantitative Concepts Age Equivalent Score	**0.411**	**0.516**	0.399	**0.443**	**0.522**

Note. Values greater than or equal to $r = .400$ are bolded. All r values greater .189 are statistically significant at $p < .001$ for $n = 299$.
WJ = Woodcock Johnson.

Table 5.

Model-estimated variance accounted for by GOLD scale scores for each direct assessment measure.

Measure	ICC	GOLD Scale Score				
		Cognitive	Literacy	Social Emotional	Language	Mathematics
Peabody Picture Vocabulary Test - Raw Score	0.234	**0.190**	**0.205**	0.080	**0.199**	**0.186**
Peabody Picture Vocabulary Test - Standard Score	0.107	0.107	0.112	0.048	0.106	0.108
Head-Toes-Knees-Shoulders (HTKS)	0.202	0.065	0.109	0.042	0.104	0.052
Pencil Tapping	0.151	0.066	0.138	0.058	0.077	0.149
Pre-LAS	0.280	**0.243**	**0.256**	0.114	**0.198**	**0.189**
WJ Letter-Word Identification Age Equivalent Score	0.221	**0.175**	**0.297**	0.102	0.063	**0.290**
WJ Understanding Directions Age Equivalent Score	0.098	0.146	**0.164**	0.089	0.156	0.105
WJ Applied Problems Age Equivalent Score	0.094	0.102	0.104	0.044	0.081	0.081
WJ Word Attack Age Equivalent Score	0.203	0.124	**0.204**	0.110	0.050	0.151
WJ Quantitative Concepts Age Equivalent Score	0.137	**0.282**	**0.397**	**0.224**	**0.232**	**0.298**

Note. Values greater than or equal to .160 are bolded and equate to an approximate r value of .400 or greater.
WJ = Woodcock Johnson.