Dialog, *17(3)*, 122-135
Copyright © 2014,
ISSN: 1930-9325

Dialog From The Field

The *Teaching Strategies GOLD*® Assessment System: Measurement Properties and Use

Diane C. Burts
Louisiana State University

Do-Hong Kim
University of North Carolina at Charlotte

This paper describes the development and use of a relatively new, widely adopted, teacher-rated authentic assessment measure and presents an overview of findings from four studies regarding its reliability and validity. The Teaching Strategies *GOLD*® is an assessment system designed to assess the development and learning of children birth - kindergarten, inclusive of children with disabilities and English language learners. Study samples were large and diverse. Children attended Head Start, private child care, and school-based programs in all areas of the United States and were representative of the nation's population of similarly aged children. Overall results indicate that the *GOLD*® yields valid and reliable inferences for the intended population and that teachers are able to use the measure to accurately assess children's development and learning across the designated age range. Limitations and implications for practice and future research are discussed.

Keywords: early childhood assessment; authentic assessment

Along with the upsurge of national, state, and local accountability guidelines and standards, child assessment has assumed particular importance for public policy discussions and educational decisions and recommendations. To guide the curriculum and assessment decisions of its various programs, Head Start identified school readiness goals for infants and toddlers (Early Head Start National Resource Center, 2012) and developed a readiness outcomes framework for children 3-5 years old (U.S. Department of Health and Human Services, 2010). As increasing numbers of children representing numerous backgrounds and experiences enroll in Head Start and other early childhood programs, scientifically-informed assessment measures are needed to ensure that

This article is based on some of the same datasets and analyses contained in a previously released technical report entitled Technical Manual for the *Teaching Strategies GOLD*® Assessment System. Partial funding for this research was provided by Teaching Strategies LCC. Opinions reflect those of the authors and do not necessarily reflect those of the granting agency.

Correspondence concerning this article should be addressed to Diane C. Burts, E-mail: dburt1@lsu.edu

all children regardless of culture, language, or disabilities are assessed fairly (Hirsh-Pasek, Kochanoff, Newcombe, & de Villiers, 2005; Snow & Van Hemel, 2008; Qi & Marley, 2009).

The purpose of this paper is to describe a relatively new authentic assessment measure, the *Teaching Strategies GOLD*® (*GOLD*®) (Heroman, Burts, Berke, & Bickart, 2010) and to present findings regarding its psychometric properties. This information is especially timely and pertinent for stakeholders given the current emphasis on assessment and the widespread use of the *GOLD*® in Head Start and other early childhood programs. Its publishers report that it is used in all states for Pre-K assessment with nearly half of the states having state-level agreements. About one-quarter of the states have state-level agreements for kindergarten assessments (J. Mosley, personal communication, September 26, 2013).

Assessment Practices

Purposeful and systematic assessment includes consideration of *why* assess, *what* should be measured, and *how* it should be measured. Assessments that are well designed, implemented effectively, and interpreted and used appropriately can inform teaching and contribute to better outcomes for all children (Snow & Van Hemel, 2008). On the other hand, poorly designed assessment measures or those that do not consider child characteristics can result in children being mislabeled or not receiving the support and services they need.

Various professional groups have provided the early childhood field with guidance on what constitutes best practices regarding assessment. For example, the National Research Council of the National Academies (Snow & Van Hemel, 2008) issued a comprehensive report on the design, implementation, and use of early childhood assessment. According to the National Association for the Education of Young Children (NAEYC) and the National Association of Early Childhood Specialists in State Departments of Education (NAECS/SDE) assessment measures should be developmentally appropriate, educationally important, and linguistically and culturally responsive (Copple & Bredekamp, 2009; NAEYC, 2009; NAEYC & NAECS/SDE, 2003; 2005). Along with these groups, the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME) point out that child assessment measures should meet established standards of reliability and validity (AERA, APA, & NCME, 1999; NAEYC & NAECS/SDE, 2003; Snow & Van Hemel, 2008).

In a policy report issued by the Society for Research in Child Development (SRCD), Hirsh-Pasek and colleagues introduce the concept of “empirical validity” (Hirsh-Pasek et al., 2005). They call for child-assessment instruments which are based on current research findings in the developmental domains with attention to those child competencies that predict later academic success. Further, they assert that assessment measures should focus on the processes (*how*) of learning. Instruments developed with process in mind portray the child’s progress along a developmental path (Hirsh-Pasek, et al., 2005). Capturing children’s emerging abilities and their performance as they engage in the active process of learning provides insights that may not be obtained otherwise and is fairer to young children who are developing and learning rapidly. During early childhood, several weeks or months can make tremendous differences in a child’s knowledge and/or abilities (Copple & Bredekamp, 2009). It is therefore important that assessments are sensitive to small changes over time (Meisels & Atkins-Burnett, 2005) and are designed to accurately capture nuanced differences in both the quantity and quality of change (Snow & Van Hemel, 2008).

Authentic Assessment

Some assessment measures rely predominately on single child responses, and the assessment information is often collected at a single point in time. Children's on-the-spot responses are influenced by various factors including their attention span, experience with assessment-like language and tasks, and culture (Atkins-Burnett, 2007). Furthermore, how they respond at any one point in the day can vary with the classroom setting and activity type (Vitiello, Booren, Downer, & Williford, 2012). Many professionals, including those in early intervention (e.g., Keilty, LaRocco, & Casell, 2009), support authentic, observation-based performance assessment of young children (e.g., Copple & Bredekamp, 2009; Gullo, 2006; Meisels, Wen, & Beachy-Quick, 2010; Moreno & Klute, 2011). Authentic assessment is ongoing, and information is collected by teachers in typical everyday situations rather than as an add-on to daily instruction (McAfee & Leong, 2011). "Careful observations of children's behavior, including how they approach tasks, create work products, and interact with others, are an ideal way for teachers to recognize and monitor children's growth on an ongoing basis" (Daniels & Perry, 2003, p. 107) and to guide classroom instruction.

In authentic assessment, information is gathered from various sources including families and others familiar with the child. Head Start has emphasized the importance of family involvement since its inception and continues to stress the import of parents as partners in their children's learning (Azar, Miller, & Stevenson, 2013). Children's development and learning are contextual and are influenced by reciprocal interactions which occur in their immediate environments (Bronfenbrenner & Morris, 2006). Close communication with families provides the teacher with valuable information (Caspe, Seltzer, Kennedy, Cappio, & DeLorenzo, 2013) and helps bridge the developmental contexts of home and school.

Teachers as Assessors

Although it seems logical that teachers who interact with children and their families on a regular basis would be in an ideal position to gather assessment data, they often struggle with the assessment process. Some researchers suggest that teachers may not provide valid and reliable information about child performance (Waterman, McDermott, Fantuzzo, & Gadsden, 2012). Teachers report that they sometimes make modifications to the measurement procedures which can affect reliability and validity (Gokiert, Noble, & Littlejohns, 2013). Teachers' evaluations of children, regardless of the type of evaluation used, are influenced by their beliefs and perceptions, education and training, life and teaching experiences, familiarity with the assessment instrument, knowledge about the assessed areas, and program settings and quality (e.g., Gallagher & Lambert, 2006; Kilday, Kinzie, Mashburn, & Wittaker, 2012; Mashburn & Henry, 2004; Mashburn, Hamre, Downer, & Pianta, 2006; Meisels et al., 2010; Phillips & Lonigan, 2010; Ready & Wright, 2011).

Authentic assessment relies on teachers' abilities to objectively observe and document what children say and do, select suitable examples and artifacts that illustrate particular knowledge and skills, and interpret and use the information appropriately (Brenneman, 2011). Teachers often have limited or no participation in training related to the assessment tools and procedures they use (Gokiert et al., 2013). "To assess appropriately and effectively, a teacher must understand at least when and how to use assessment; how a child's development can affect the assessment process; and the intersections among assessment, program evaluation, and curriculum and teaching" (Gullo, 2006, p. 138). Given the complexity of the task, some teachers may be unprepared to meet this challenge, particularly with the vast diversity of children represented in programs today. For example, assessing the skills of second language learners can be difficult for teachers (Rodrigues & Guiberson, 2011), especially if they do not understand children's cultural backgrounds and home languages (Espinosa, 2010) or the process of learning a second language (Tabors, 2008).

The aforementioned factors make it particularly important that assessment measures are not only psychometrically sound and empirically valid, but that they are also "teacher-friendly." Teachers must be able to interpret the language and ratings used in the tool without ambiguity (Goldstein & McCoach, 2011; Meisels et al., 2010). The following sections provide an overview of the *Teaching Strategies GOLD*® and summaries of the research findings from studies related to its psychometric properties.

METHODS

Instrument

Overview. The *Teaching Strategies GOLD*® (Heroman et al., 2010) is an authentic, observation-based teacher rating system designed to assess the continuing development and learning of children birth through kindergarten, inclusive of second language learners and children with disabilities. The measure is intended to assist teachers in planning and individualizing instruction and in monitoring and communicating child progress with families and other stakeholders. Although it is closely tied to *The Creative Curriculum*® *for Infants, Toddlers, & Twos* and *The Creative Curriculum*® *System for Preschool*, its developers indicate that the measure can be used in programs using the curricula and those that use other curricula (Teaching Strategies LLC, n.d.).

The *GOLD*® has 25 research-based objectives which are operationalized into 51 rating scale items organized into the areas of social–emotional (9 items), physical (5 items), language (8 items), cognitive (10 items), literacy (12 items), and mathematics (7 items). A research section precedes each objective and provides its foundation by summarizing major research findings, including those predictive of future success.

Development. Development of the *GOLD*® occurred over several years with consideration of the guidelines provided by professional groups (e.g., AERA, APA, & NCME, 1999; Hirsh-Pasek, et al., 2005; NAEYC & NAECS/SDE, 2003; Snow & Van Hemel, 2008). Solicited feedback from teachers, administrators, consultants, and Teaching Strategies, LLC professional-development staff was incorporated into the measure. Two pilot studies conducted with diverse populations further refined the measure. National experts (e.g., development, content areas, special needs, ELLs) provided content review relating to the objectives, item importance, developmental progressions, sample behaviors, age ranges, research, cultural sensitivity, organization, clarity, etc. Final assessment items resulted from the comments 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 current research and professional literature including literature identifying the knowledge, skills, and behaviors most predictive of school success.

An initial study of the *Teaching Strategies GOLD*® was conducted with 290 children who ranged in age from 2.5 months to 35 months. Children attended programs in 20 different centers located in all regions of the United States. Results indicated adequate internal consistency reliability and validity of the measure (Kim & Smith, 2010).

Assessment Process. The *GOLD*® objectives help focus the assessment process as teachers gather information through observations, conversations, artifacts, etc. during daily activities. For example, during a project/study on houses the teacher may notice a child constructing different types of houses. While observing,the teacher may document progress related to the child's interactions with peers, engagement in conversations, task persistence, and use of fingers and hands. Documentation may be gathered using various means such as audio or video recordings, photographs, or observational notes and 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 *GOLD*®. 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." They include examples of readily observable behaviors which help teachers assess child progress toward the objective. Levels 1, 3, 5, 7, and 9 are the "In-between" levels and do not include examples. They are used to indicate that the child's skills in the area are emerging but aren't established and may need support (e.g., physical, visual, verbal, gestures, modeling). These additional steps in the progressions better portray the nuances in children's development and learning than yes/no ratings or fewer steps are able to capture. Color-coded bands which overlap indicate the typical age and/or grade-level (i.e., kindergarten) ranges for each item. Some bands are longer or shorter than others and indicate the uneven and overlapping nature of development and learning (Copple & Bredekamp, 2009). Suggested teaching strategies, inclusive of children of various ages, cultures, language abilities, and disabilities, are provided for each objective.

STUDIES

Participants

The psychometric properties of any new assessment instrument should be rigorously examined and made available to stakeholders. When establishing validity and reliability, large samples representative of the children with whom the measure will be used, are necessary. This gives teachers and administrators confidence that the assessment can be successfully used with children in different parts of the country; in diverse instructional settings; and with children of different backgrounds, races, ethnicities, and special needs. The four studies described in this paper represent the information provided from two national norm samples. Samples were broad and diverse but the sizes varied depending on the particular research questions.

The population (n=111,059) included ratings made by children's classroom teachers using the *GOLD*® for the fall 2010 checkpoint. A norm sample (n_1=10,963) was created that matched the 2009 U.S. Census Bureau estimates for children ages birth to 71 months with respect to ethnic subgroups, gender, census region, and state. The population was divided into three-month age bands, for a total of 24 age bands ranging from 0-2 months to 69-71 months. Using a similar procedure, a second, longitudinal norm sample (n_2=33,612) was created from among all children who were rated at all three checkpoints during the 2010-2011 academic year. Detailed descriptions of the sampling procedures are reported in separate manuscripts (Kim, Lambert, & Burts, 2013; Kim, Lambert, & Burts, in press; Lambert, Kim, & Burts, 2014a; 2014b).

Children were enrolled in programs at 2,525 different Head Start, private childcare, and school-based sites in 48 states and the District of Columbia. Most of the participating programs used *The Creative Curriculum® for Infants, Toddlers, and Twos* and/or *The Creative Curriculum® for Preschool* (Teaching Strategies LCC, n.d.) and had been using the developmental continua previously developed by Teaching Strategies LCC (2001; 2005; 2006). Teachers were transitioning to the *GOLD*® assessment system. A total of 4,580 teachers gathered child data using the *GOLD*®. Prior to the studies, teachers participated in a two-day training focusing on an overview of the measure and exploration of the objectives and child progressions. They watched video clips, participated in large-group discussions, evaluated a portfolio, completed family conference forms, and practiced uploading documentation samples and entering observation notes and progress checkpoint data online

Study 1

Purpose, Methods, and Findings. Study 1 was conducted to explore evidence for the reliability and validity of the information provided by the *GOLD*® with children birth through kindergarten. The study addressed the (a) factorial structure of the measure, (b) indexes of reliability, and (c) inter-rater reliability. Both norm samples (n_1=10,963; n_2=33,612) provided the data for the study. The first step was to examine the factorial structure of the measure using confirmatory factor analysis (CFA). A six-factor model was selected because of its basis in developmental theory and its correspondence with the structure of the instrument. Results of CFA based on the first norm sample indicated that the six-factor model fit the data reasonably well (Standardized Root Mean Square Residual [SRMR] = .033, Comparative Fit Index [CFI] = .932, and Root Mean Square Error of Approximation [RMSEA] = .066). All factor loadings were generally large and statistically significant ($p < .001$). The correlations between the six scales were also large (.786 - .958) and statistically significant ($p < .001$).

Using the second norm sample data, longitudinal measurement invariance CFA was conducted to examine whether the scores obtained when teachers use the *Teaching Strategies GOLD*® measured the intended constructs equivalently across time. The results indicated the *Teaching Strategies GOLD*® construct was equivalent across time.

Reliability was evaluated using person separation index, person reliability, item separation index, and item reliability provided by Rasch analyses. The person separation index indicates how well the instrument can differentiate persons (i.e., children) 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. Cronbach's alpha was used to measure internal consistency. Rasch analyses indicated that the six scales (i.e., social-emotional, cognitive, etc.) appear to yield scores that are highly reliable. The Cronbach's alpha reliability coefficients indicated very high internal consistency (.94 - .97).

Correlations between the ratings of a master trainer and ratings of teachers (n=557) who were current users of the system provided inter-rater reliability information. The master trainer rated 18 children (age 5 to 71 months) on all items in the measure. The teachers looked at video clips of the same children and provided their ratings across all items in the assessment. Teachers provided 2,558 separate child assessments and rated an average of 4.59 children. Each teacher rated only those children in the age-group or class/grade with whom she/he worked. The correlations between the master trainer and raters were high. All but one was above .90; it was above .80. No data were collected on how well the teachers applied the *GOLD*® training information when they assessed the children in their actual classrooms.

Discussion. Findings from Study 1 (Lambert et al., 2014a) supported the construct-related validity of the measure suggesting that the *GOLD*® measures six separate domains as intended by its developers. Results of longitudinal invariance CFA indicated the constructs were equivalent across time implying that the interpretations of changes in children's development and learning obtained from the measure are valid. Not particularly surprising, correlations between factors were high. Although distinct developmental and learning domains are identified in the literature, these areas are overlapping and interrelated (e.g., Copple & Bredekamp, 2009; Hirsh-Pasek et al., 2005; Snow & Van Hemel, 2008). While teachers must be able to differentiate between various areas so they can take appropriate, early action they must also be cognizant of the interwoven nature of domains and how each area influences and is influenced by other areas. This holistic view supports children's overall well-being and success (Copple & Bredekamp, 2009) and is less likely to promote concentration in one area at the expense of others.

Reliability is an extremely important concern when considering teacher as ratings. Inter-rater reliability between a master trainer and teachers was high. Reliability coefficients for all three checkpoints were also high. In agreement with an initial study of the measure with children birth through age two (Kim & Smith, 2010), these results suggest that teachers can reliably assess the development and learning of children birth through 71 months using the *GOLD*®.

Study 2

Purpose, Methods, and Findings. Another important consideration, and one especially germane to Head Start, is whether the measurement is invariant across subgroups of children. Study 2 looked specifically at whether the *GOLD*® is valid for children with disabilities and those for whom English is not their first language. Data from three-, four-, and five -year-olds with complete item responses were used in the study. Assessment information collected for the fall (n=79,324), winter (n=132,693), and spring (n=50,558) checkpoint periods were analyzed according to each child's primary language or disability status, forming three groups: (a) children with disabilities were compared to those without disabilities; (b) ELLs were compared to children whose primary language is English (non-ELLs); and (c) Spanish-speaking ELLs were compared to non-ELLs. Differential Item Functioning (DIF) analysis was used to determine if any items within the *GOLD*® were operating differently for the subgroups under study. Findings showed that the majority of items in the *GOLD*® displayed little or no DIF.

Several language items revealed DIF at one checkpoint, but not at others. However, the item, “uses conventional grammar” was consistently identified as having DIF for all subgroups at all three checkpoint periods.

Discussion. When establishing an instrument’s validity, it is important to note that acknowledged group differences do not constitute biased ratings (DeVeliis, 2003). Instrument and item bias have specific statistical definitions (Clauser & Mazor, 1998) based on findings of differential item or test functioning (DIF/DTF) after statistically controlling for child ability. Findings from Study 2 indicated that in general, teachers’ ratings were similar for children of similar abilities, regardless of their subgroup membership as indicated by the absence of DIF. Although DIF was not detected for most items in the *GOLD*®, it was consistently detected for the item, “uses conventional grammar” (Kim et al., 2013). As such, the presence of DIF indicates the possibility of bias for that item, but it does not necessarily indicate bias. For example, a secondary dimension being measured (auxiliary) may cause benign DIF (Douglas, Roussos, & Stout, 1996). Items in the *GOLD*® that showed DIF, particularly the item related to grammar usage, should be carefully monitored in the future. Further, both pre-service education and in-service training should emphasize how ELLs and children with disabilities demonstrate their language abilities (Rothstein-Fisch, Trumbull, & Garcia, 2009).

Study 3

Purpose, Methods, and Findings. One important aspect in determining the validity of an observational assessment is the variability attributed to the child versus other factors. This study addressed: (a) child characteristics associated with teacher ratings (i.e., age, gender, disability status, and English language status); (b) classroom composition characteristics associated with teacher ratings (i.e., class mean age; and percent boys, disabilities, and ELLs); and (c) the variability between raters when controlling for child and classroom characteristics. Three-level growth curve modeling was used to analyze the data. A sample of 21,592 children was selected from among children rated during all three rating periods using the online version of the *GOLD*®. Scale scores were created for each developmental domain using interval level Rasch rating scale ability estimates (Bond & Fox, 2007). Findings indicated that teacher ratings were associated in anticipated directions for child characteristics and for classroom characteristics. Children with disabilities began the year behind their typically developing peers and grew at slower rates throughout the year. Girls showed an advantage in some areas over boys. ELLs were rated lower at the beginning of the year but showed somewhat faster growth rates than native English-speakers. Approximately 16% and 25% of the variance in scale scores was accounted for by unmeasured differences between classrooms and teachers, including rater effects.

Discussion. The findings from Study 3 suggest that the *GOLD*® showed sensitivity to age differences and to growth over time. As expected, older children had higher scores at all checkpoints than younger children. Consistent with findings from other research, children with disabilities started behind their non-disabled peers and grew at slower rates over the year (Gallagher & Lambert, 2006; Goldstein, 2004). Boys began the year lower and grew slower than girls in all areas except mathematics, partially supporting findings of some researchers (Penner &

Paret, 2008; Robinson & Lubienski, 2011) and contrasting with others (Klein, Adi-Japha, & Hakak-Benizri, 2010; Matthews, Ponitz, & Morrison, 2009). Similar to findings of other studies (e.g., Downer, Lopez, Grimm, Hamagami, Pianta, & Howes, 2012; Reardon & Galindo, 2009; Yesil-Dagli, 2011), ELLs in general were rated lower at the beginning of the year than English speakers, and in some cases grew at faster rates than non-ELLs. This finding could indicate individual differences among children, some of whom acquire a second language easily while others need more time. An alternative hypothesis is an initial “mismatch” (Ray, Bowman, & Brownell, 2006) between teachers and ELLs or perhaps that teachers did not fully understand the English-language acquisition process (Espinosa, 2010; Tabors, 2008). As children became more proficient in English or as teachers became familiar with the children, teachers may have provided more accurate ratings.

Another important factor in determining instrument validity is the amount of error variance, or how much variability can be ascribed to the child as opposed to other factors. Teacher–based observational assessment is more subjective than direct assessments (Cabell, Justice, Zucker, & Kilday, (2009) thereby creating the possibility for greater variability (Kilday et al., 2012). High unexplained variance of as much as 50% has been reported in some studies of teachers’ global ratings of children (Mashburn & Henry, 2004). The error variance reported in Study 3 (Lambert et al., 2014b) was considerably lower than that noted in some studies (e.g., Kilday et al., 2012; Mashburn & Henry, 2004). The design of the *GOLD*® (e.g., research-based objectives, multiple examples, additional scale points, behavioral anchors along the progressions) may address some of the problems related to teacher-based ratings found in other studies.

Study 4

Purpose, Methods, and Findings. The purpose of Study 4 (Kim et al., in press) was to develop interval level scale scores that could be used to track development and learning across the entire age range of the *GOLD*®. The study focused on (a) the examination of dimensionality, (b) rating scale effectiveness, (c) hierarchy of item difficulties, and (d) the relationship of developmental scale scores to child age. Data from the first norm sample (n_1=10,963) were analyzed using the Rasch Rating Scale Model (Andrich, 1978).

Results indicated that each subscale in the measure satisfies the Rasch model for unidimensionality (i.e., items measure only one underlying latent construct). The 10-category rating structure (i.e., “not yet” to level 9) functioned effectively, with few exceptions; ratings at the very lowest and highest ends of the scale were less reliable and in-between ratings were less distinctive. In general, items formed theoretically expected hierarchies (e.g., items which are less difficult for children were rated by teachers as lower item difficulty). Moderately high correlations of developmental scale scores with child age were found.

Discussion. Results from Study 4 provide additional supporting evidence for the construct validity of the *GOLD*®. Further, results suggest that teachers can make valid ratings of the developmental progress of children across the measured age range. Contrary to suggestions by some authorities that teachers may respond less accurately to item anchor scales with more anchor points and subtle progressions (McDermott et al., 2011), overall, study teachers were able to distinguish among the 10 levels of progressions with ratings for the lowest and highest levels being somewhat less reliable. The smaller sample sizes at both ends of the progressions likely contributed to this finding. In addition, the Rasch category probability curves indicated "In-between" levels seemed redundant with adjacent categories, indicating that teachers were less clear about the "in between" levels. These levels are important steps in the progressions (particularly for children with special needs) as children demonstrate that their knowledge, skills, and behaviors in the area are emerging. Teachers must be able to recognize when children's abilities are beginning so they can scaffold their development and learning by providing the appropriate support (Early, Iruka, Ritchie, Barbarin, Winn, Crawford et al., 2010). Because the in-between levels do not include examples, future training on the *GOLD*® should focus on these levels in depth to ensure that teachers fully understand them and can adequately provide the child support needed.

SUMMARY AND IMPLICATIONS

Early childhood classroom teachers are increasingly called upon to provide assessment information and to make important decisions about children based upon the assessment results. It is therefore critical that the assessment instruments they use are not only "teacher friendly" (Goldstein & McCoach, 2011) but that they are also empirically valid (Hirsh-Pasek et al., 2005) and psychometrically sound (AERA, APA, & NCME, 1999; NAEYC & NAECS/SDE, 2003; Snow & Van Hemel, 2008). The studies reported in this paper add to the research literature on teacher-rated authentic assessment by exploring the validity and reliability of the recently developed *Teaching Strategies GOLD*® assessment system.

Overall, findings from the four studies reported in this paper suggest that the *GOLD*® provides a viable teacher-rated assessment option for assessing areas of development and learning for children birth through kindergarten representing diverse backgrounds. The process used in its development and its design support the principles of empirical validity (Hirsh-Pasek, et al., 2005). Accumulated evidence indicates that the measure is psychometrically sound, culturally and linguistically responsive, and sensitive to children with disabilities (Kim et al., 2013; Kim et al., in press; Lambert et al., 2014a, 2014b). As such, the measure can be an effective tool for monitoring child progress and helping teachers plan and individualize instruction as they gather information about children during daily classroom activities and settings. It can also help teachers determine which children might benefit from individualized instruction or further evaluation. Moreover, knowing the developmental sequence through which children typically progress can help teachers plan for children with a range of abilities and support families in better understanding development and learning trajectories (Early et al., 2010).

Despite questions by some researchers concerning teachers' abilities to accurately rate children's abilities (Waterman et al., 2012), findings imply that teachers were able to use the *GOLD*® to make valid ratings of the developmental progress of children, allowing for children's development and learning to be tracked longitudinally. Using one system which can be used to assess program children ages birth through kindergarten is beneficial for tracking children's development and learning longitudinally (Snow & Van Hemel, 2008). It can also promote program assessment continuity as children move from one classroom to the next because teachers are already familiar with the assessment system.

Limitations and Implications for Future Research

Although the studies reported in this paper represent assessment data from large and diverse groups of children in Head Start and other types of programs in all parts of the United States, several limitations must be noted. Sample sizes at the extreme ends of the progressions (i.e., 0-2 months, 3-5 months, and 69-71 months) were smaller which may have influenced some of the research findings by precluding some types of analyses with these age groups. Future studies should aim to include larger samples of children in these age ranges. In addition, certain child information (e.g., type of disability) and teacher characteristics (e.g., educational level/training, years of teaching experience, and cultural background) were not available to researchers. Their absence somewhat limited the conclusions drawn from the results. Given the heterogeneity of children with disabilities, it will be important in future studies to explore particular child disabilities (e.g., those influencing language) in relationship to measured items. There is also a need for studies that examine a range of teacher characteristics (e.g., education, assessment training, and program settings and quality) and their associations with how teachers assess children using the *GOLD*®. This would help clarify if the between rater variance found in Study 3 is associated with teacher characteristics or with other variables; understanding this relationship could help reduce their potential impact on teacher ratings. Although reliability with a master trainer was established, information regarding how the teachers actually applied material covered in the training was not available to the researchers. This information would strengthen conclusions about reliability and teachers' abilities to use the *GOLD*® with fidelity.

In sum, the findings from the studies reported in this paper add much to the empirical research on authentic assessment and to the utility of the *Teaching Strategies GOLD*® in particular. The widespread use of the *GOLD*® in Head Start and other programs makes it critical that its adequacy be evaluated. As with any assessment tool, appraisal of the psychometric properties of the measure will need to continue as long as it is being used. Although there were only a few items that showed less valid and reliable results, they should be vigilantly examined by instrument developers and carefully monitored in replication studies. In addition, future studies should examine the relationships between teacher ratings using the *GOLD*® and the children's scores on other measures collected by outside assessors. These quantitative results along with qualitative studies of teachers (e.g., how they use the tool to inform instruction) could provide helpful information for classroom assessment and insights that may prove useful in any future revisions of the *GOLD*®.

REFERENCES

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrich, D. (1978). Application of a psychometric model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*, 581-594.

Atkins-Burnett, S. (2007). Measuring children's progress from preschool through third grade. Mathematica Policy Research, Inc. Retrieved from http://www.mathematica-mpr.com/publications/PDFs/measchildprogress.pdf

Azar, S. T., Miller, E. A., & Stevenson, M. T. (2013). Promoting engagement and involvement of parents with cognitive challenges: Suggestions for Head Start programs. *Dialog, 16*(1), 216-235.

Bond, T., G., & Fox, C., M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Bordignon, C., M., & Lam, T., C. M. (2004). The early assessment conundrum: Lessons from the past, implications for the future. *Psychology in the Schools, 41*, 737-749.

Brenneman, K. (2011). Assessment for preschool science learning and learning environments. *Early Childhood Research and Practice, 13*(1). Retrieved from http://ecrp.uiuc.edu/v13n1/brenneman.html

Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology. Vol. 1: Theoretical models of human development* (6th ed., pp. 793-828). New York, NY: Wiley.

Cabell, S. Q., Justice, L. M., Zucker, T. A., & Kilday, C. R. (2009). Validity of teacher report for assessing the emergent literacy skills of at-risk preschoolers. *Language and Speech & Hearing Services in Schools*, *40*, 161-173.

Caspe, M., Seltzer, A., Kennedy, J. L., Cappio, M., & DeLorenzo, C. (2013). Engaging families in the child assessment process. *Young Children, 68*(3), 8-14.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31-44.

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.

Daniels, D. H., & Perry, K. E. (2003). “Learner-centered" according to children. *Theory into Practice, 42*(2), 102-108.

DeVeliis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks, CA: SAGE.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement*, *33*(4), 465-484.

Downer, J. T., Lopez, M. L., Grimm, K. J., Hamagami, A. , Pianta, R. C., & Howes, C. (2012). Observations of teacher-child interactions in classrooms serving Latinos and dual language learners: Applicability of the Classroom Assessment Scoring System in diverse settings. *Early Childhood Research Quarterly*, *27*, 21-32.

Early, D. M., Iruka, I.U., Ritchie, S., Barbarin, O.A., Winn, D-M. Crawford, G., et al. (2010). How do pre-kindergarteners spend their time? Gender, ethnicity, and income as predictors of experiences in pre-kindergarten classrooms. *Early Childhood Research Quarterly, 25*, 177-193.

Early Head Start National Resource Center (2012). *School readiness goals for infants and toddlers*. Washington, DC: Author. Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/ttasystem/ehsncr/Early%20Head%20Start/early-learning/curriculum/school-readiness-goals-infants-toddlers.pdf

Espinosa, L. (2010). *Getting it right for children from diverse backgrounds: Applying research to improve practice*. Upper Saddle River, NJ: Pearson Education.

Gallagher, P. A., & Lambert, R. G., (2006). Classroom quality, concentration of children with special needs, and child outcomes in Head Start. *Exceptional Children 73*(1), 31-52.

Gokiert, R., Noble, L., & Littlejohns, L. B. (2013). Directions for professional development: Increasing knowledge of early childhood measurement. *Dialog, 16*(3), 1-20.

Goldstein, J., & McCoach, B. (2011).The starting line: Developing a structure for teacher ratings of students skills at kindergarten entry. *Early Childhood Research and Practice*, 13(2). Retrieved from http://ecrp.uiuc.edu/v13n2/goldstein.html

Goldstein, P.. (2004). Helping young children with special needs develop vocabulary. *Early Education Journal, 32*, 1-43.

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.

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

Hirsh-Pasek, K., Kochanoff, A., Newcombe, N. S., & deVilliers, J. (2005). Using scientific knowledge to inform preschool assessment: Making the case for “empirical validity. *Social Policy Report, 14*(1), 3-19.

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.

Kilday, C.R., Kinzie, M. B., Mashburn, A. J., & Wittaker, J.V. (2012). Accuracy of teacher judgments of preschoolers’ math skills. *Journal of Psychoeducational Assessment 30,* 148-159.

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*, 574-595.

Kim, D-H., Lambert, R. G., & Burts, D, C. (in press). Validating a developmental scale for young children: Applicability of the *Teaching Strategies GOLD*® assessment system. *Journal of Applied Measurement*.

Kim, D-H., & Smith, J. D. (2010). Evaluation of two observational assessment systems for children's development and learning. *NHSA Dialog, 13*, 253-267.

Klein, P. S., Adi-Japha, E., & Hakak-Benizri, S. (2010). Mathematical thinking of kindergarten boysandgirls: Similar achievement, different contributing processes. *Educational Studies in mathematics, 73*, 233-246.

Lambert, R. G., Kim, D-H., & Burts, D. C. (2014a). The measurement properties of the *Teaching Strategies GOLD*® assessment system. Manuscript submitted for publication.

Lambert, R. G., Kim, D-H., & Burts, D. C. (2014b). Using teacher ratings to track the growth and development of young children using the *Teaching Strategies GOLD*® assessment system. *Journal of Psychoeducational Assessment, 32*(1), 27-39.

Mashburn, A. J., Hamre, B. K., Downer, J. T., & Pianta, R. C. (2006). Teacher and classroom characteristics associated with teachers' ratings of prekindergartners' relationships and behaviors. *Journal of Psychoeducational Assessment, 24*, 367-380.

Mashburn, A. J., & Henry, G. T. (2004). Assessing school readiness: Validity and bias in preschool and kindergarten teachers' ratings. *Educational Measurement Issues and Practices, 23*(4), 16-30.

Matthews, J. S., Ponitz, C. C., & Morrison, F. J. (2009). Early gender differences in self- regulation and academic achievement. *Journal of Educational Psychology, 101*, 689-704.

McAfee, O., & Leong, D. J. (2011). *Assessing and guiding young children's development and learning* (5th ed.).Upper Saddle River, NJ: Pearson.

McDermott, P. A., Fantuzzo, J.W., Warley, H. P., Waterman, C., Angelo, L. E., Gadsden, V. L. et al. (2011). Multidimensionality of teachers' graded responses for preschoolers' stylistic learning behavior: The Learning-to-Learn Scales. *Educational Psychological Measurement 71*(1), 148-169.

Meisels, S. J., & Atkins-Burnett, S. (2005). *Developmental screening in early childhood: A guide*(5th ed.). Washington, DC: National Association for the Education of Young Children.

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.

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 (2009). *Where we stand on assessing young English language learners*. Retrieved from www.naeyc.org/positionstatements/

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 from www.naeyc.org/dap

NAEYC & NAECS/SDE (2005). *Screening and assessment of young English-language learners: Supplement to the NAEYC and NAECS/SDE joint position statement on early childhood curriculum, assessment, and program evaluation*. Retrieved from www.naeyc.org/dap

Penner, A. M., & Paret, M. (2008). Gender differences in mathematics achievement: Exploring the early grades and the extremes. *Social Science Research, 37*, 239-253.

Phillips, B. M., & Lonigan, C.J. (2010). Child and informant influences on behavioral ratings of preschool children. *Psychology in the schools, 47*, 374-390.

Qi, C. H., & Marley, S. C. (2009). Differential item functioning analysis of the *Preschool Language Scale-4* between English-speaking Hispanic and European American children from low-income families. *Topics in Early Childhood Special Education, 29*(3), 171-180.

Ray, A., Bowman, B., & Brownell, J. O. (2006). Teacher-child relationships, socio-emotional development, and school achievement. In B. Bowman & E. K. Moore (Eds.), *School readiness and social-emotional development: Perspectives on cultural diversity* (pp. 7-22). Washington, DC: National Black Child Development Institute, Inc.

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal, 48*, 335-360.

Reardon, S. F., & Galindo, C.. (2009). The Hispanic-White achievement gap in math and reading in the elementary grades. *American Educational Research Journal, 46*, 853-891.

Robinson, J. P., & Lubienski, S. T. (2011). The development of gender gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal, 48*, 268-302.

Rodriguez, B. L., & Guiberson, M. (2011). Using a teacher rating scale of language and literacy skills with preschool children of English-speaking, Spanish-speaking, and bilingual backgrounds. *Early Childhood Education Journal, 39*, 303-311.

Rothstein-Fisch, C., Trumbull, E., & Garcia, S. G. (2009). Making the implicit explicit: Supporting teachers to bridge cultures. *Early Childhood Research Quarterly, 24*, 474-486.

Snow, C. E., & Van Hemel, S. B. (Eds.) (2008). *Early childhood assessment: Why, what, and how? Report of the National Research Council of the National Academies*. Washington, DC: National Academies Press. Retrieved from http://www.nap.edu/catalog/12446.html

Tabors, P. (2008). *One child, two languages: A guide for early childhood educators of children learning English as a second language* (2nd ed.). Baltimore: Paul H. Brookes.

Teaching Strategies LCC. (n. d.). *Teaching Strategies for early childhood: Curriculum*. Retrieved from www.teachingstrategies.com

Teaching Strategies LCC. (2001). *Creative Curriculum Developmental Continuum for Ages 3-5,*Washington, DC: Author.

Teaching Strategies LCC. (2005). *Expanded Forerunners of the Creative Curriculum Developmental Continuum for Ages 3-5*, Washington, DC: Author.

Teaching Strategies LCC. (2006). *Creative Curriculum Developmental Continuum for Infants, Toddlers, and Twos.* Washington, DC: Author.

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 years old*. Washington, DC: Author.

Vitiello, V. E., Booren, L. M., Downer, J. T., & Williford, A. (2012). Variation in children's classroom engagement throughout a day in preschool: Relations to classroom and child factors. *Early Childhood Research Quarterly, 27*, 210-220.

Waterman, C., McDermott, P. A., Fantuzzo, J. W., & Gadsden, V. L. (2012). The matter of assessor variance in early childhood education – Or whose score is it anyway*? Early Childhood Research Quarterly, 27,* 46-54.

Yesil-Dagli, U. (2011). Predicting ELL students' beginning first grade English oral reading fluency from initial kindergarten vocabulary, letter naming, and phonological awareness skills. *Early Childhood Research Quarterly, 26*, 15-29.