Outcomes Study

LEVEL OF EVIDENCE
Gold Standard

Teaching Strategies GOLD® Assessment System

Technical Summary

Summary Findings of a Study Conducted by
The Center for Educational Measurement and Evaluation
The University of North Carolina at Charlotte

Teaching Strategies GOLD® Assessment System

Technical Summary

Introduction

Selecting a Meaningful Assessment Instrument

When selecting an assessment instrument to administer to children, the most important considerations are the *validity* and *reliability* of the measure. Validity refers to *what* the assessment tool measures and *how well* it does so. Reliability refers to the consistency of scores obtained for the same children when reexamined with the same assessment instrument on different occasions, with different sets of equivalent items, or under other variable assessment conditions.

To ensure that *Teaching Strategies GOLD*® is both valid and reliable, The Center for Educational Measurement and Evaluation (CEME), The University of North Carolina at Charlotte, conducted extensive research with thousands of children and teachers. This document is a summary of the results obtained from that research.

Teaching Strategies GOLD® Overview

Teaching Strategies GOLD® is an authentic observation-based assessment system for children from birth through kindergarten. The system may be implemented with any developmentally appropriate curriculum. It blends ongoing observational assessment for all areas of development and learning with performance tasks for selected predictors of school success in the areas of literacy and numeracy. *Teaching Strategies GOLD*® can be used to assess all children, including English-language learners, children with disabilities, and children who demonstrate competencies beyond typical developmental expectations.

Using *Teaching Strategies GOLD*®

The primary purpose of *Teaching Strategies GOLD*® is to document children's learning over time, inform instruction, and facilitate communication with families and other stakeholders. It is important to remember that *Teaching Strategies GOLD*® is not intended as a screening or diagnostic measure, an achievement test, or a program-evaluation tool.

Objectives for Development and Learning

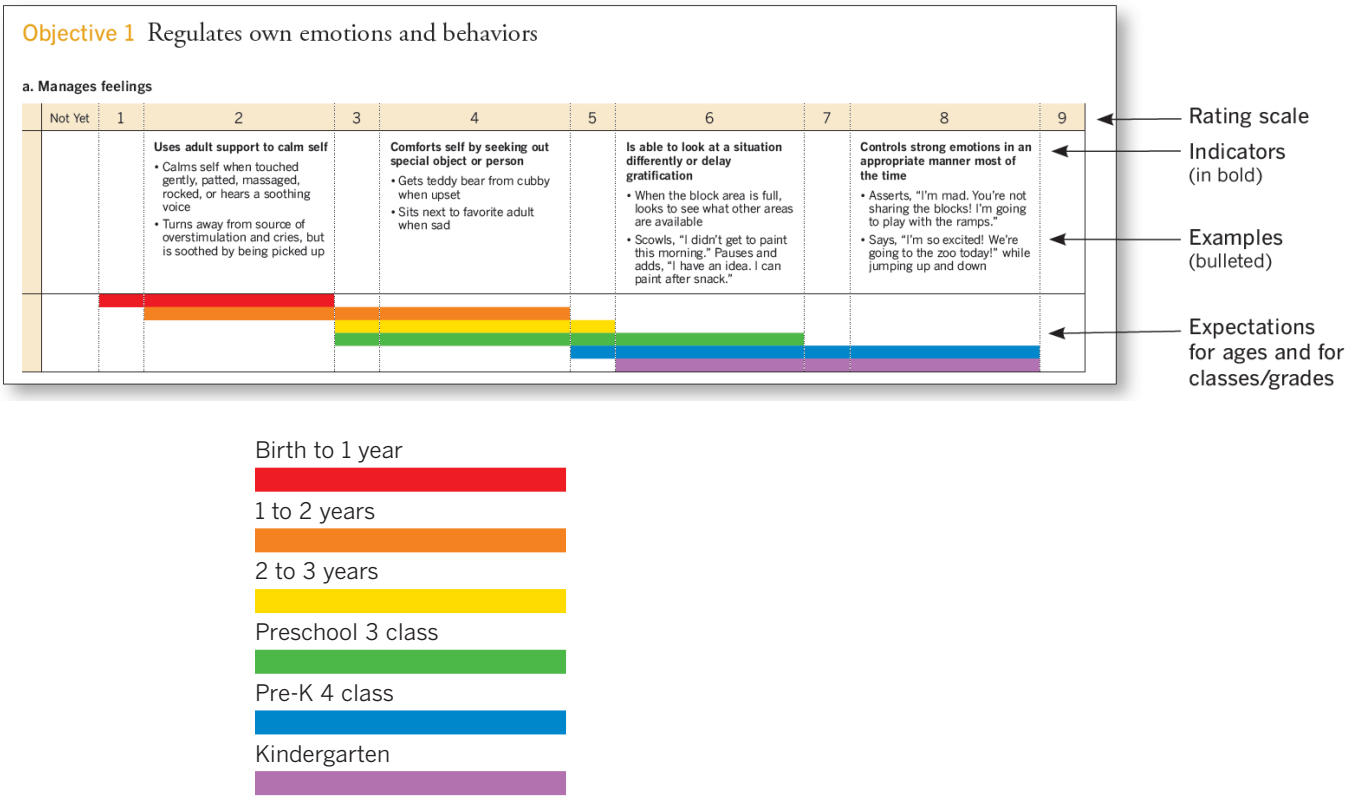
Teaching Strategies GOLD® enables educators to focus on and measure the knowledge, skills, and behaviors most predictive of school success. The tool has a total of 38 objectives. Two objectives are related specifically to English language acquisition, and the other 36 objectives are organized into nine areas of development and content-area learning. The areas are

- Social–Emotional
- Physical
- Language
- Cognitive
- Literacy
- Mathematics
- Science and Technology
- The Arts

Teaching Strategies Gold® Progressions of Development and Learning

Teaching Strategies Gold® presents progressions of development and learning for objectives in the areas of social–emotional, physical, language, and cognitive development and in the content areas of literacy, mathematics, and English-language acquisition. Indicators and examples enable tool administrators to rate children's knowledge, skills, and behaviors on a 10-point scale of "Not Yet" to level 9. Furthermore, with the exception of those for English language acquisition, the progressions use colored bands to show widely held expectations for various ages (birth–1 year, 1–2 years, and 2–3 years) and for various classes/grades (preschool 3, pre-K 4, and kindergarten). At a glance, these colored bands show educators and families which skills and behaviors are typical for children of a particular age or class/grade. The bands help teachers manage the complexity of young children's development, which *Teaching Strategies Gold*® recognizes as being uneven and rapidly changing. They also help teachers and families understand that expectations for a particular age or class often overlap expectations for another.

Here is the progression for Objective 1, Dimension a:



The Norm Sample

Determining the Sample

When determining the validity and reliability of an early childhood assessment instrument, it is important to identify a large sample of children who are representative of the nation's population of similarly aged children. Doing so allows teachers and administrators to assume that the instrument will be used equally effectively with children from all parts of the country; children in all types of instructional settings; and children with different backgrounds, races, ethnicities, and special needs.

CEME determined the norm sample from a total of 111,059 children rated by using *Teaching Strategies GOLD®*. The total population was divided into 3-month age bands, for a total of 24 age bands ranging from 0–2 months to 69–71 months. Teachers answered questions about each child's background, race, and ethnicity that were identical to those employed by the U.S. Census Bureau. The goal was to represent each of the twenty-four 3-month age bands with 500 randomly selected children. This sampling procedure was used to match the U.S. Census Bureau 2009 estimates for children ages birth to 5 years, 11 months with respect to seven ethnic subgroups.

Final Sample

The final sample used to evaluate the validity and reliability of *Teaching Strategies GOLD®* retained a total of 10,963 children. This extremely diverse group of children received educational services in 618 different programs at 2,525 different early childhood centers located across the United States. These programs included Head Start, private child care, and school-based sites. Forty-eight states and the District of Columbia were represented in the final sample. A total of 4,580 teachers was selected as raters to administer *Teaching Strategies GOLD®*. Overall, the final sample used in this research was large, broad, and highly representative of young children in the United States.

Construct Validity

Construct validity refers to whether the assessment instrument measures the theoretical constructs (e.g., knowledge, skills, or behaviors) that it is intended to measure. To determine whether *Teaching Strategies GOLD®* is a valid tool for measuring early childhood development and learning, several analyses were conducted.

Factors Measured by *Teaching Strategies GOLD®*

The first step was to confirm the areas of development that *Teaching Strategies GOLD®* is intended to measure. Researchers examined a six-factor model that corresponded to the design of the instrument. This model evaluated each assessment item's "fit" within one of six areas: social–emotional, physical, language, cognitive, literacy, and mathematics. Statistically, the study's goal was to find a Root Mean Square Error of Approximation (RSMEA) value of <.06, a Standardized Root Mean Square Residual (SRMR) value of < .08, and a Comparative Fit Index (CFI) value of at least .90. The overall results supported the six-factor design of *Teaching Strategies GOLD®* with a RMSEA = .066, a SRMR = .033, and a CFI = .931. All of these analyses were statistically significant at $p < .001$, demonstrating that the assessment instrument reliably measures those six factors of child development (social–emotional, physical, language, cognitive, literacy, and math).

Scale and Item Analysis

Researchers further conducted an analysis known as Rasch scaling to determine that the six areas of *Teaching Strategies GOLD®*, and the items within those areas, measure one and only one factor (e.g., social–emotional but *not* language development.) This is also referred to as *unidimensionality*. For each of the six areas (social–emotional, physical, language, cognitive, literacy, and mathematics), the results of the analysis indicated that they are unidimensional, meaning they are distinct from one another and acceptably measure only one factor within the overall assessment. Furthermore, with the exception of one literacy item and one mathematics item, all individual objectives and dimensions within each area of *Teaching Strategies GOLD®* are distinct and measure only one of the six areas.

Rating Scale Effectiveness

The items in *Teaching Strategies GOLD®* are measured on a 10-point scale from level 0 to level 9. Researchers evaluated the rating process for each of the six scales to determine whether teachers were administering the instrument in the way it was intended. Statistical analysis should ideally demonstrate that the average performance on the various scales strictly advanced as the individual ratings advanced, which was the case for the social–emotional, physical, and cognitive scales. For the language, literacy, and mathematics scales, two of the possible ratings on the scale (e.g., 0 vs. 1 and 7 vs. 8) overlapped, indicating that the descriptions of those particular ratings might have been somewhat redundant and therefore challenging for teachers to discriminate between when evaluating children.

Item Difficulty

Finally, researchers evaluated the specific items within the six factors of *Teaching Strategies GOLD®* to determine whether they progress in difficulty as expected for typically developing children. Results confirmed that the six factors (social–emotional, physical, language, cognitive, literacy, and mathematics), or scales, consisted of items that increased in difficulty and align with accepted developmental milestones. According to the CEME researchers, the developers of *Teaching Strategies GOLD®* were "very successful in creating measures that offer a developmental pathway of sequential milestones that agree with developmental theory."

Reliability

Several analyses were conducted to determine whether *Teaching Strategies GOLD®* is a reliable measure of development and learning. These included person and item reliabilities, internal consistency reliability, and interrater reliability.

Person and Item Reliabilities

High person and/or item reliability means that there is a high probability of replicating the instrument's results. Specifically, person reliability estimates the likelihood of children's performing the same across other items measuring the same constructs of child development as those measured by *Teaching Strategies GOLD®*. Similarly, item reliability estimates the likelihood that the instrument's items would follow the same developmental progression if administered to another sample of children with similar abilities. Person and item reliabilities

of .8 and higher are considered strong indicators of reliability. Across the six scales of *Teaching Strategies GOLD®*, person reliabilities ranged from .95 to .98, while item reliabilities were .99 for all six scales. These values indicate very high person and item reliability for *Teaching Strategies GOLD®*.

Internal Consistency Reliability

Internal consistency reliability refers to the consistency of children's responses to all items within each area of the instrument. The more homogeneous the domain measured, the higher the internal consistency reliability should be. Researchers measured the internal consistency of the items within each area of *Teaching Strategies GOLD®*. They determined internal consistency reliability estimates ranging from .957 for the physical scale to .980 for the cognitive scale. These values represent extremely high internal consistency reliability.

Interrater Reliability

Interrater reliability refers to the consistency of scores obtained when two different people administer the same instrument to the same child. If the tool is reliable, the results should be the same (or nearly the same) regardless of the user. Researchers conducted an interrater reliability study by examining the correlations between the ratings of a *Teaching Strategies GOLD®* master teacher/trainer and the ratings of teachers to whom the assessment system is new. This study was conducted by first having a master teacher/trainer rate the skills of 18 children on all items of the instrument. Next, a sample of 557 teachers examined video clips of the same children and provided their ratings for all assessment items. Each teacher rated the skills of only those children who matched the age-group he or she worked with, meaning that no teacher rated all 18 children. Researchers determined the correlations at the area level (e.g., physical, cognitive, language, etc.) between the teacher ratings and the master teacher/trainer ratings. Correlations were very high, with all but one being above .90 and the lowest correlation still being high at .80. The highest level of agreement between the master teacher/trainer and the new teachers was found in the literacy scale. This is very strong evidence of interrater reliability for *Teaching Strategies GOLD®*.

Scale Scores and Age Bands

Scale scores are generally considered more reliable and meaningful than raw scores when analyzing assessment data. Researchers determined scale scores for *Teaching Strategies GOLD®* by using strategies common in both educational and psychological testing. Children's ability estimates were rescaled to conform to a normal distribution with a mean of 500 and standard deviation of 100. Scores three or more standard deviations below the mean were given a value of 200, while values three or more standard deviations above the mean were given a value of 800. Data was analyzed by separating children into 3-month age bands based on their age in months at the time of the first assessment checkpoint in October 2010.

Results indicate that the mean for each scale score is appropriately occurring around age 36 months, which is the middle age range for which *Teaching Strategies GOLD®* is intended. Scale scores correlate moderately strongly with age, suggesting that teachers are generally giving higher scores to older children and lower scores to younger children. Since *Teaching Strategies GOLD®* is meant to measure progress across skills that follow a developmental progression, these results are positive and promising. Furthermore, mean scores for the age bands increase with age at a steady pace. This finding indicates that *Teaching Strategies GOLD®* can be used to track and monitor the developmental progress of children from year to year.

Differential Item Analysis

Assessment instruments should ideally be valid and reliable with all populations of children, including those with disabilities and those for whom English is not a home language. Researchers used differential item analysis to determine whether any items of *Teaching Strategies GOLD®* were operating differently for different populations of children. Three age-groups (3-, 4-, and 5-year-olds) were selected for this study. Data was analyzed according to each child's primary language and disability status, forming three groups of interest: children with disabilities, English-language learners (ELLs), and Spanish-speaking children. There is strong evidence that the items in *Teaching Strategies GOLD®* are operating the same way for different groups of children, meaning that the assessment instrument is equally valid and reliable for children with special needs and for those whose home language is not English.

Conclusion

The *Teaching Strategies GOLD®* assessment system yields highly valid and reliable results. The results of the current research strongly validates that teachers are able to use *Teaching Strategies GOLD®* to make valid ratings of the developmental progress of children across the intended age range from birth through kindergarten. Future analysis will focus on the variance in the ratings that can be attributed to child age, within teacher variability and between teacher variability. Additional evidence of concurrent validity will be released in fall 2011.

TeachingStrategies.com | 800.637.3652