March 2021

NeMTSS Research Brief

Teaching Strategies GOLD and a Balanced Assessment System

Abril Rangel-Pacheco, M.A. & Amanda Witte, Ph.D.

NEBRASKA CENTER FOR RESEARCH ON CHILDREN, YOUTH, FAMILIES & SCHOOLS

NeMTSS Implementation Support Team, University of Nebraska–Lincoln is funded by the Nebraska Department of Education (USDE IDEA, Part B-611 Grant). Contents do not necessarily represent the policy of NDE or USDE, and no endorsement should be assumed. Developed by UNL-CYFS MTSS Implementation Support Team.

Teaching Strategies GOLD and a Balanced Assessment System: An NeMTSS Research Brief

Key Points:

- Child assessment is an important component of high-quality early childhood education programs. It is a vital tool for understanding and supporting children's development, and it is essential to document program effectiveness (Epstein et al., 2001).

- The Teaching Strategies GOLD (TS GOLD) assessment system has been found to have adequate construct and convergent validity, as well as empirically supported structure and reliability measures. However, there has been less support regarding the assessment's discriminant validity. This indicates a possible limitation in the TS GOLD to accurately capture within-child variability across learning domains and in its ability to differentiate readiness skills between children of the same classroom (Russo et al., 2019; Miller-Bains, 2017).

- Early childhood education (ECE) programs should clearly delineate their intended purposes for a readiness assessment and consider investigating and documenting the appropriateness of the TS GOLD assessment system for their various uses (Lambert et al., 2015a; Russo et al., 2019)

A Balanced Assessment System and NeMTSS

A system intentionally using formative, interim, and summative data to inform instruction and program development, monitor progress, and evaluate student learning for all content areas and grade level is the heart of a Multi-Tiered System of Support (MTSS; Underwood, 2020; Nebraska Department of Education, 2018; Nebraska Department of Education, 2017). Data from a strategically balanced assessment system can provide meaningful and interpretable information for stakeholders at all levels in the educational system (Educational Testing Service, 2018). Due to the importance of evidence-based assessments in an NeMTSS framework (Nebraska Department of Education, 2018), assessment measures and instruments must be both reliable and valid to optimize interpretation.

Assessment in Early Childhood Education

Assessment in early childhood has garnered growing attention over the years, most notably as a means to (a) evaluate program efficacy and justify the appropriation of educational funds and (b) evaluate student learning and school readiness (Epstein et al., 2001). With data from these assessments being used in "high-stakes" decisions, it is critical to understand common challenges associated with conducting valid and reliable assessments in early childhood.

One major challenge to assessment in early childhood education is that the younger the child, the more difficult it can be to obtain valid assessment measures (Epstein et al., 2001). Development in the early years is rapid, episodic, and strongly influenced by experience and environmental supports (Qiu et al., 2021; Epstein et al., 2001). Additionally, performance on an assessment can be affected by children's emotional states and the conditions of the

assessments, which creates more sources of measurement error and negatively affects measurement validity and reliability (Epstein et al., 2001).

Other issues common with early childhood assessments are the alignment of the assessment content with state/institutional learning standards, rater differences, adequate floor/ceiling items to capture the profile of children from different ability levels, and administration time challenges (Ackerman, 2018).

Teaching Strategies GOLD and Early Childhood Assessment

One of the most widely used performance-based, observational measures in early childhood education programs is the TS GOLD (Heroman et al., 2010) (Russo et al., 2019). The TS GOLD assessment is a multidimensional, performance-based, observational assessment where teachers observe children's skills during typical instruction across nine broad areas of development: literacy, mathematics, language, social-emotional, cognitive, physical, science and technology, social studies, and arts (Heroman et al., 2010). It measures the knowledge, skills, and attributes that are "most predictive of school success" (Heroman et al., 2010), and such information allows early education teachers to monitor children's ongoing development and learning; to individualize educational expectations and design developmentally appropriate curricula; to identify children who might need further evaluation, screening, or special assistance; and to communicate with parents about their development (Heroman et al., 2010; Russo et al., 2019). In the initial edition, the assessment system covered ages birth through kindergarten (TS GOLD B–K). The assessment system was recently extended to birth through third grade in its new edition (TS GOLD B-3rd) (Lambert, 2017) and details surrounding the differences between the versions can be found in (Qiu, et al., 2021).

Although the TS GOLD assessment system *was not* designed as a teacher/program evaluation tool, an achievement test, or a screening tool, it assesses children's competencies that are predictive of school readiness (Heroman et al., 2010), and has been widely adopted for accountability purposes across the nation (Miller-Bains et al., 2017). Given its growing use in states across the country, it is vital to understand the evidence supporting its psychometric properties.

Reliability and Validity of Teaching Strategies GOLD

There are several studies that have investigated the psychometric properties of the TS GOLD, and while results have generally supported the reliability and validity of TS GOLD system, some studies have had mixed findings.

Russo et al., (2019) investigated the TS GOLD B-K assessment's convergent validity, discriminant validity, and intraclass correlations (ICC). The study found adequate convergent validity relative to direct assessments in the fall, spring, and over the course of the preschool year. In other words, associations between children's readiness skills as measured by TS GOLD and direct assessments within a skill area (e.g., literacy) were positive and modest to moderate in strength. Convergent validity was also supported by Miller-Bains (2017) which found the assessment to have strong associations with independent direct assessments of the same constructs (e.g., Woodcock-Johnson III Tests of Achievement). However, both Russo et al., (2019) and Miller-Bains (2017) found limited evidence for the assessment's discriminant validity.

Russo et al., (2019) conducted intraclass correlations (ICC) to examine discriminant validity and found that proportions of classroom-level variance for the TS GOLD domains were markedly larger than those of the direct assessments in the fall, spring, and across the preschool year. The differences in ICCs between TS GOLD and the direct assessments showed

that when children within a classroom were assessed by independent data collectors using direct assessments, they looked much less similar to one another than when they were assessed by teachers using the TS GOLD. Miller-Bains (2017) found a similar pattern, in which their ICC differences indicated that students' scores in each learning construct were much more similar within a given classroom compared to those produced by the direct assessments.

Russo et al., (2019) found that across analyses and time points, teachers' assessments of children's readiness skills using TS GOLD lacked precision in discriminating between both children within a classroom and skills within a child. The authors found that overall, when compared to direct assessments, TS GOLD showed limited ability to differentiate among children's readiness skills and between children within a classroom both in the fall, spring and over the course of the year. A similar pattern emerged in Miller-Bains (2017), which found that teachers tended to rate individual students more similarly across all learning constructs despite empirical evidence of more substantial variation across domains when skills were measured via direct assessment. Miller-Bains (2017) posited that readiness ratings produced using TS GOLD may be more influenced by factors that are separate from a child's actual skills compared to results obtained from direct assessments. Taken together, Russo et al., (2019) and Miller-Bains (2017) suggest that this evidence may indicate a limitation in the TS GOLD to accurately capture within-child variability across learning domains and in its ability to differentiate readiness skills between children in a given classroom.

Qiu et al., (2021) investigated the construct validity and reliability of the newer, extended version of the assessment: TS GOLD B-3rd grade. The authors found evidence supporting the construct validity of the assessment via strong associations between item responses and the specific areas of child development the assessment was designed to measure. High reliability estimates (above 0.95) were collected for each TS GOLD B-3rd domains (Qiu et al., 2021), suggesting TS GOLD B-3rd domain scores to be a reliable measure of young children's competence in the six development and learning areas. According to the authors, the findings from this study are in line with previous studies validating the measurement structure of TS GOLD B–K (Lambert, 2017a; 2017b). Lambert (2017a; 2017b) reported evidence of unidimensionality for each of the six domains and reported multiple indices of reliability including Cronbach's alpha, person reliability, and item reliability. For Cronbach's alpha, coefficients of the six domains ranged from 0.96 to 0.99. Evidence supporting the reliability and construct validity of the TS GOLD B-K was also noted in Lambert, Kim and Burts (2014).
Similar claims were made in other studies (Kim et al., 2013, Lambert et al., 2015a, Lambert et al., 2015b). These studies also obtained high internal consistency reliability estimates across domains (Cronbach's alpha above 0.90) and high reliability estimates. According to Qiu et al., (2021) these earlier studies found no evidence of statistical bias (differential item functioning) and verified consistency of the measurement structure over time and across subgroups of children.

Qiu et al., (2021) believes the TS GOLD B-3rd to be a psychometrically adequate instrument to the extent of its internal structure, of the six development domains (i.e., social-emotional, physical, language, cognitive, literacy, and mathematics) for children from birth through pre-kindergarten, however, not without some limitations (for a detailed account, see Qiu et al., 2021). Qiu and colleagues report a need to investigate the newer TS GOLD B-3rd version of the assessment and specifically, a need for more validation studies to investigate the issue of discriminant validity raised in Russo et al., (2019) and Miller-Bains (2017).

To summarize, empirical evidence has generally supported the TS GOLD B-K and the subsequent TS GOLD B-3rd as measures with adequate psychometric properties, however, there have been mixed results regarding discriminant validity, which may indicate a limited ability to accurately differentiate among student's readiness skills and between children within a classroom. Russo et al., (2019) theorized that this may be due to global similarities in children of the same classroom and/or rater and environmental effects influencing the scores within

children of the same classroom. However, this has yet to be investigated. This lack of discrimination between domains within children in a classroom can be a concern when assessment performance is used for screening purposes (Qiu et al., 2019). Because this issue has not been examined yet, it may be beneficial for educators and stakeholders to interpret the results with caution and be mindful of the limitations of the TS GOLD assessment system and early childhood assessments in general.

Implications for Practitioners

Given that assessment results in early childhood can vary in validity and reliability, it may be beneficial to interpret test/assessment scores as part of a broader assessment that includes components like observations, portfolios, work samples, or ratings from teachers and/or parents (Epstein et al., 2001). Including both strengths and areas of needed support may better characterize a child's school readiness profile better than one comprehensive assessment (Russo et al., 2019).

Assessments or instruments with inadequate validity and reliability make it difficult for stakeholders to gain insight into a range of skills for individual students and aggregated to the class, school, district, and state levels (Russo et al., 2019). In light of this issue, a balanced approach to collecting data as part of early childhood assessment should be mindful of common issues related to early childhood assessment such as, fluctuations in child development, administration time and timeline constraints, test alignment, and test construction (Epstein et al., 2001).

ECE programs should clearly delineate their intended purposes for a readiness assessment (i.e., individualization of supports for children, tracking of program level outcomes over time, targeting professional development or curriculum) (Russo et al., 2019), and it is recommended that all current and prospective users of the TS GOLD assessment system investigate and document the appropriateness of this measurement system for their various uses (Lambert et al., 2015a).

References

Ackerman, D. J. (2018). Real world compromises: Policy and practice impacts of kindergarten entry assessment-related validity and reliability challenges. *ETS Research Report Series*, *2018*(1), 1-35. https://doi.org/10.1002/ets2.12201

Educational Testing Service (2018). *Understanding Balanced Assessment Systems*. https://www.ets.org/s/k12/pdf/ets-k-12-understanding-measurement-white-paper.pdf

Epstein A. S., et al. (2004). *Preschool Assessment: A Guide to Developing a Balanced Approach.* National Institute for Early Education Research. https://nieer.org/wp-content/uploads/2012/03/7.pdf

Heroman, C., Burts, D. C., Berke, K., & Bickart, T. S. (2010). *Teaching strategies GOLD® objectives for development & learning: Birth through kindergarten.* Washington, D.C: Teaching Strategies.

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 & Development*, *24*(4), 574-595. https://doi.org/10.1080/10409289.2012.701500

Lambert, D.H. Kim, D.C. (2015b) *Using teaching strategies GOLD to assess kindergarten readiness and track growth and development*. Technical report. Center for Educational Measurement and Evaluation, University of North Carolina Charlotte, Charlotte, N.C https://ceme.uncc.edu/sites/ceme.uncc.edu/files/media/Lambert,%20Kim,%20Burts.pdf

Lambert, R (2017b) *Evidence supporting the use of GOLD® with kindergarten children.* Technical report. Center for Educational Measurement and Evaluation, University of North Carolina Charlotte, Charlotte, N.C https://ceme.uncc.edu/sites/ceme.uncc.edu/files/media/GOLD%20Technical%20Report%20-%202017%20with%20cover.pdf

Lambert, R. (2017a). *Technical manual for the Teaching Strategies GOLD® assessment system: Birth through third grade edition*. Technical Report. Charlotte, NC: Center for Educational Measurement and Evaluation, University of North Carolina Charlotte. https://teachingstrategies.com/wp-content/uploads/2018/05/CEMETR-2017-02-Lambert_0.pdf

Lambert, R. G., Kim, D. H., & Burts, D. C. (2014). 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. https://doi.org/10.1177/0734282913485214

Lambert, R. G., Kim, D. H., & Burts, D. C. (2015a). The measurement properties of the Teaching Strategies GOLD® assessment system. *Early Childhood Research Quarterly*, *33*, 49-63, https://doi.org/10.1016/j.ecresq.2015.05.004

Miller-Bains, K. L., Russo, J. M., Williford, A. P., DeCoster, J., & Cottone, E. A. (2017). Examining the validity of a multidimensional performance-based assessment at kindergarten entry. *AERA Open*, *3*(2), https://doi.org/10.1177/2332858417706969

Nebraska Department of Education (2017). *S12- Position Statement on Assessment of Student Learning*. Nebraska State Board of Education Position Statements https://www.education.ne.gov/policyreference/position-statement-on-assessment-of-student-learning/

Nebraska Department of Education. (2018). NeMTSS Framework. https://nemtss.unl.edu/

Qiu, Y., Leite, W. L., Rodgers, M. K., & Hagler, N. (2021). Construct validation of an innovative observational child assessment system: Teaching Strategies GOLD® birth through third grade edition. *Early Childhood Research Quarterly*, *56*, 41-51. https://doi.org/10.1016/j.ecresq.2021.02.005

Russo, J. M., Williford, A. P., Markowitz, A. J., Vitiello, V. E., & Bassok, D. (2019). Examining the validity of a widely-used school readiness assessment: Implications for teachers and early childhood programs. *Early Childhood Research Quarterly*, *48*, 14-25. https://doi.org/10.1016/j.ecresq.2019.02.003

Underwood, S. (2020) *How to build a balanced assessment system*. NWEA. https://www.nwea.org/blog/2020/how-to-build-a-balanced-assessment-system/#:~:text=A%20balanced%20assessment%20system%20intentionally,emphasis%20placed%20on%20formative%20assessment.

Recommended Citation:

Rangel-Pacheco, A. & Witte, A. L. (2020). *Teaching Strategies GOLD and a Balanced Assessment System: An NeMTSS Research Brief.* Nebraska Multi-tiered System of Support (NeMTSS).

Authorship Information:

Abril Rangel-Pacheco, M.A.
School Psychology Doctoral Student
Graduate Research Assistant
Nebraska Center for Research on Children, Youth, Families and Schools
University of Nebraska–Lincoln
arangel-pacheco2@huskers.unl.edu