*AERA Open*
*April-June 2017, Vol. 3, No. 2, pp. 1–16*
*DOI: 10.1177/2332858417706969*


# Examining the Validity of a Multidimensional Performance-Based Assessment at Kindergarten Entry

**Katherine L. Miller-Bains**
**Jaclyn M. Russo**
**Amanda P. Williford**
**Jamie DeCoster**
**Elizabeth A. Cottone**
*University of Virginia*

*The present study explored the utility of a widely used performance-based assessment of children's readiness skills as a kindergarten entry assessment. In a sample of 520 kindergarten students across 52 classrooms, we compared students' school readiness skills as assessed by teachers using Teaching Strategies GOLD (TS GOLD) to direct assessments administered by independent data collectors. We found evidence of the concurrent validity of TS GOLD. However, the measure demonstrated weaknesses in its ability to differentiate readiness skills. Specifically, the highly correlated latent constructs and intraclass correlations associated with TS GOLD indicated that teachers were less likely to differentiate readiness skills among children within the same classroom relative to direct assessments. The findings are discussed in terms of assessing children's school readiness at scale. As the number of programs, districts, and states that require school readiness assessments increases, so does the need to better understand the information that can be inferred from particular methods of assessment.*

Keywords: *school readiness, kindergarten entry assessment, early childhood, validity, reliability, authentic assessment, direct assessment, observational assessment, performance-based assessment, formative assessment*

With increased focus on improving and expanding children's early educational experiences, policymakers and practitioners across the country have a vested interest in ensuring all students enter kindergarten ready to learn. Over the past decade, the education community has recognized the importance of school readiness, or the set of skills children possess at school entry that are linked to later academic and social success (K. Snow, 2006). As research on school readiness has expanded, so has its definition to encompass multiple domains of learning, including language, literacy, general knowledge, approaches to learning, physical health, and social and emotional development (K. Snow, 2006; Zill, Collins, West, & Hausken, 1995). A combination of child and environmental characteristics influences the development of school readiness, resulting in natural variation in children's exhibited skills (Blair & Raver, 2015; Brooks-Gunn & Markman, 2005; Magnuson, Meyers, Ruhm, & Waldfogel, 2004). However, pronounced differences in readiness across students within the same school, district, or classroom can be substantial and consequential, particularly when comparing students from different socioeconomic backgrounds (Isaacs, 2012; Rimm-Kaufman, Pianta, & Cox, 2000).

The large learning disparities that emerge during early childhood contribute to gaps that are present when children enter kindergarten (Magnuson et al., 2004; Lee & Burkam, 2002; Pratt, McClelland, Swanson, & Lipscomb, 2016; Rimm-Kaufman et al., 2000). These early differences can persist and become wider as students progress through the educational system, leading to a gamut of negative outcomes in later school and life (Belsky & MacKinnon, 1994; Duncan et. al., 2007; Galindo & Sonnenschein, 2015; Hamre & Pianta, 2001; Sadowski, 2006). However, early targeted intervention has been shown to improve the performance of children who initially lag behind their more advantaged peers (Conroy & Brown, 2004; Dunlap, Johnson, & Robbins, 1990; S. Ramey & Ramey, 2004; Walker et al., 1998) and has been most effective at influencing long-term outcomes (Belfield, Nores, Barnett, & Schweinhart, 2006; F. Campbell & Ramey, 1994, 1995; Ferrer et al., 2015; C. Ramey & Campbell, 1991; Schweinhart, 1993).

The success of early interventions suggests a need to accurately identify children's learning needs, which can in turn inform education practice and policy. Early childhood assessments contribute a crucial piece of feedback as these measures provide information about both students' learning needs as they enter kindergarten as well as the types of experiences children have prior to school entry (K. Snow, 2011). Establishing statewide assessment systems can therefore

serve the needs of classroom teachers and instructional specialists by providing data on children's current skill levels, which can be used to inform instructional practices in the classroom while also enabling division leaders and educational policymakers to examine the learning gaps that might be addressed through programmatic preschool and early elementary interventions.

Statewide kindergarten entry assessments (KEAs) have been used for the aforementioned purposes by providing information regarding young children's incoming skills across a range of early learning domains across a large population of students. Most KEAs are intended to inform instruction and identify students who fall below developmental expectations in order to intervene in areas of need (Shields, Cook, & Greller, 2016). Whereas some states have focused their KEAs on a single domain (such as Idaho's Reading Indicator and the Phonological Awareness Literacy Screening in Wisconsin), many departments of education have incorporated multidimensional measures that assess students' skills across more than one area of readiness. Regardless of the dimensions of readiness, these entry assessments come in multiple formats, including checklists/rating scales, performance-based assessments, and direct assessments (Connors-Tadros, 2014).

The majority of states with mandated entry assessments have adopted performance-based measures—also known as observation-based, authentic, naturalistic, or work sampling assessments—for use as their statewide KEAs. Broadly, performance-based measures permit the teacher to rate a child's skills on a set of items after a period of observing and documenting the student's functioning, usually as it occurs naturally in the classroom context (K. Snow, 2011). Scholars and educators have asserted that these types of assessments offer several advantages over more standardized tools. Most commonly, educational practitioners have argued that performance-based assessments are more developmentally appropriate for young children (Macy & Bagnato, 2010; K. Snow, 2011), enable the measurement of skills and behaviors as they occur naturally rather than in an artificial or unfamiliar setting (Bagnato & Macy, 2010; Dennis, Rueter, & Simpson, 2013), are better at informing intervention/instruction than alternative forms of assessment (Dennis et al., 2013; Macy & Bagnato, 2010; Wiggins, 1990), and can pose less imposition on teachers, students, and classroom time as skills are assessed during regular instruction (McAfee & Leong, 2011; K. Snow, 2011).

However, the advantages of performance-based assessments must be considered along with the intended use of the subsequent ratings of students' skills. The psychometric properties (e.g., reliability and validity) of any assessment provide insight into a measure's utility for varied purposes as well as the extent to which practitioners can trust the results produced by a given measure (C. Snow & Van Hemel, 2010). The higher the demonstrated consistency—or reliability—of an assessment, the more comparable scores will be across the classrooms, raters, and contexts in which they are administered. This is particularly useful for large-scale KEAs as many educational stakeholders are able to gain insight into the same set of skills for individual students and aggregated to the class, school, district, and state levels (C. Snow & Van Hemel, 2008; K. Snow, 2011). Generally, the more standardized the assessment protocol and procedures, the higher the reliability of the measure will be, as standardization limits the likelihood that extraneous and subjective influences, such as a rater's preconceptions or administration inconsistencies, will factor into students' scores (C. Snow & Van Hemel, 2008; Waterman, McDermott, Fantuzzo, & Gadsden, 2012).

Regardless of a measure's consistency, an assessment tool is valuable only if it is valid—or captures necessary and meaningful information regarding the domain it is supposed to measure. Validity considers whether the discrete skills measured by an instrument align in expected ways to the larger learning construct in the target population. In some instances, higher reliability can come at the sacrifice of greater validity, as greater standardization can narrow the scope of an assessment, which can in turn limit the measure's ability to capture relevant information about students' abilities in a specific learning domain (C. Snow & Van Hemel, 2008). Therefore, when making comparisons across children or classrooms, early educators must attempt to balance the breadth of data obtained and flexibility of administration with safeguards against inconsistencies across administrations that may obscure students' relative abilities (Pavelski-Pyle, 2002; K. Snow, 2011).

In this study, we examined the utility and psychometric properties of one such multidimensional performance-based assessment used by many states as a KEA: Teaching Strategies GOLD (TS GOLD; Heroman, Burts, Berke, & Bickart, 2010). More specifically, we focused on evidence of validity within the KEA context and compared the performance of this teacher-administered performance-based measurement tool to direct assessments of the same skills conducted by independent data collectors.

## KEAs

### *Current Use of Kindergarten Entry Assessments*

In response to the need for accurate assessments of children's readiness skills, state-level departments of education across the country have instituted kindergarten entry measures intended to serve a variety of purposes, from evaluating early educational opportunities to informing teachers' instruction and identifying students who may not be ready for kindergarten (Shields et al., 2016). Although the use of school-level entry assessments has remained fairly consistent over the past decade (Little, Cohen-Vogel, & Curran, 2016; Shields et al., 2016), the use of statewide multidimensional assessments of kindergarten readiness has grown and

continues to trend upward (Maxwell, Scott-Little, Pruette, & Taylor, 2013; U.S. Department of Education, 2013). Although only seven states collected kindergarten entry data during the 2009–2010 school year, more than 25 states had some form of KEA in place as of 2012 (Connors-Tadros, 2014; National Center on Quality Teaching and Learning, 2013). By 2014, 33 states applied for federal funding through Race to the Top–Early Learning Challenge (RTT-ELC), which required the implementation of a KEA. The response to RTT-ELC indicates that the number of states systematically measuring these skills is poised to increase as national policies continue to encourage the use of KEAs (Connors-Tadros, 2014).

### *Assessment Methods Used for KEAs*

In accordance with the National Research Council's recommendations for early childhood assessments (C. Snow & Van Hemel, 2008), RTT-ELC suggested that KEAs should minimally cover five areas of learning (language and literacy, cognition and general knowledge, approaches toward learning, physical well-being, and motor development) and that the intended use of the assessment should be clearly delineated and aligned with the chosen assessment. In practice, the purpose of KEAs is often multifaceted, with needs in one or more of four categories: (a) identifying children's readiness status relative to a predetermined developmental benchmark, (b) informing teachers' instructional practices aimed at narrowing the school readiness gap at kindergarten entry, (c) developing and targeting resources for practitioners and students, and (d) informing early childhood initiatives and programs. States have chosen different assessment approaches and are forced to strike a balance between stakeholder needs, such as the timeliness and depth of data reporting, with practical limitations, such as cost and administration time. Some states have relied on off-the-shelf assessments, whereas others have modified preexisting measures. Others still have developed their own KEAs (Connors-Tadros, 2014; Siddens, Hubbell, & Otto, 2013).

Although there has been some variation in the structure and content of state KEAs, the majority of states utilizing entry measures have opted for classroom-embedded performance-based assessments, also known as authentic or observation-based assessments. In a recent search of documentation available on state websites, at least 10 states were either using or considering the use of TS GOLD as their primary KEA as of the 2015–2016 school year (Colorado Department of Education [DOE], 2016; Korobkin, 2012; Louisiana DOE, n.d.; Massachusetts Executive Office of Education, 2016; Michigan DOE, 2016; Minnesota DOE, n.d.; Nevada Early Childhood Advisory Council, 2016; New Jersey DOE, n.d.; State of Alabama DOE, 2015; State of Washington, 2015). Fourteen additional states were implementing another performance-based measure or an assessment with a large observational component as their school entry assessment during the 2015–2016 school year (Alaska Department of Education and Early Development, 2016; Arkansas DOE, n.d.; California DOE, 2016; Connecticut DOE, n.d.; Florida Office of Early Learning, 2014; Illinois State Board of Education, 2016; Maine DOE, 2015; Maryland State Board of Education, 2015; Missouri Department of Elementary and Secondary Education, n.d.; New Mexico Public Education Department, n.d.; Ohio DOE, 2016; Public Schools of North Carolina, n.d.; Vermont Agency of Education, 2016; Wyoming DOE, 2016).

Many of the performance-based assessments used by states, such as the Work Sampling System (Meisels, Jablon, Marsden, Dichtelmiller, & Dorfman, 2001), the Desired Results Developmental Profile–Kindergarten (California DOE, 2016), and TS GOLD (Heroman et al., 2010), follow a similar format and align closely in terms of the skills and domains assessed (Lambert, Kim, & Burts, 2015; Teaching Strategies, 2010). For each of these measures, teachers are asked to collect data from multiple sources, including student work, observations, and classroom tasks. Based on the information collected for each student, teachers rate children's skills across multiple domains, such as approaches to learning, language and literacy development, and mathematics ability. Generally, the skills are described and a rating scale is used to identify a student's ability level across different readiness domains. All three of the aforementioned assessments provide information about how a student is doing in each area relative to research-based developmental expectations with the intention of helping teachers plan for students' learning needs in a comprehensive manner (California DOE, 2016; Heroman et al., 2010; Meisels et al., 2001).

Despite the relative newness of performance-based assessments, most states have opted for this format over direct assessments. Although direct assessments have been used by clinicians and other educational specialists to measure young children's early academic skills for a longer period of time (Atkins-Burnett, 2007; Dennis et al., 2013; C. Snow & Van Hemel, 2008), the need for specialized training often makes them more costly and/or complicated to administer than performance-based alternatives. Within the direct assessment format, the administrator interacts with each student individually and asks the child to respond to series of highly standardized prompts or complete a sequence of tasks, and the administrator then records the child's responses to each item. The specific skills assessed by each of the items within a given measure are intended to represent a sampling of the abilities included in the overall construct of interest (Koretz, 2002). Direct assessments place more restrictions on the ways in which the child is presented with tasks and how the assessor can prompt or respond to a student. This is done to ensure that students have very similar experiences across contexts/

administrators, increase the comparability of scores, and improve the objectivity of the measure and ability to compare scores across children.

Authentic assessments offer several potential advantages over direct assessments that make them appealing, particularly in the early childhood setting. First and foremost, pencil-and-paper tests traditionally administered in upper-elementary grades and higher are neither feasible nor desirable in the kindergarten setting (C. Snow & Van Hemel, 2008). Although one-on-one direct assessments can circumvent issues with young children's ability to independently read and write, they are often time-consuming and sometimes require a student to focus on the assessment task for longer than is developmentally appropriate (Atkins-Burnett, 2007; Macy & Bagnato, 2010). Additionally, when administered by the teacher, the lengthiness of direct assessments can translate into a loss of instructional time (McAfee & Leong, 2011). Alternatively, if the assessment is completed by someone unfamiliar to the student, it can cause the child to become nervous or uncomfortable, compromising the validity of the results and introducing measurement error (Atkins-Burnett, 2007; C. Snow & Van Hemel, 2008). Performance-based assessments, on the other hand, can often be completed outside of class time and, in the case of teacher-administered authentic assessment, are less intrusive to both students and teachers as information is collected over the course of regular instruction (K. Snow, 2011). Last, these types of assessments are purported to capture more complete and relevant information about students' skills as teachers are able to compile and consider multiple sources of the child's performance in context (Atkins-Burnett, 2007). Because of these advantages, the performance-based measures that were once used in conjunction with a specific curriculum, such as TS GOLD, have evolved to become stand-alone entry assessments.

However, some have expressed concern over the potential influence of factors that are unrelated to students' actual skills on the accuracy of performance-based teacher ratings (Cabell, Justice, Zucker, & Kilday, 2009; Hoyt, 2000; Kilday, Kinzie, Mashburn, & Whittaker, 2012; Mashburn, Hamre, Downer, & Pianta, 2006; Sudkamp, Kaiser, & Jens, 2012). Raters may systematically differ in the scores that they assign to students by being too lenient, too severe, or less discriminating across skills or students or by rating all students as average (Engelhard, 1994). Such tendencies introduce score variation that is attributable to rater characteristics rather than students' demonstrated abilities, what Waterman and colleagues (2012) refer to as "assessor variance." Providers of performance-based assessments have attempted to safeguard against rater effects by providing reliability training or checks (Cash, Hamre, Pianta, & Myers, 2011) or carefully defining scoring rubrics to reduce subjectivity (Atkins-Burnett, 2007). However, these efforts are not consistent. For instance, although TS GOLD provides interrater reliability training, it does not require that teachers administering the assessment pass or complete the training process (Teaching Strategies, 2011). Despite these concerns, multiple studies have examined the psychometric functioning of performance-based assessments and have found promising evidence for both the reliability and validity of these types measures in samples of young children (Burts & Kim, 2014; Halle, Zaslow, Wessel, Moodie, & Darling-Churchill, 2011; Karelitz, Parrish, Yamada, & Wilson, 2010; Kim & Smith, 2010; Meisels, Liaw, Dorfman, & Nelson, 1995; Lambert, Kim, & Burts, 2014, 2015; Soderberg et al., 2013; Teaching Strategies, 2013a, 2013b). Below, we have included an overview of prior research on the psychometric properties of the measure utilized in this study.

Several studies have examined the reliability and validity of TS GOLD in samples of preschool (Burts & Kim, 2014; Kim & Smith, 2010; Lambert et al., 2014, 2015; Teaching Strategies, 2013a, 2013b) and kindergarten students (Lambert et al., 2015; Soderberg et al., 2013). Focusing on the two most relevant studies to the current research, Lambert and colleagues (2015) used a sample of preschool students, and Soderberg and colleagues' (2013) examined TS GOLD as a KEA in Washington state. In terms of validity, both studies focused largely on the convergent (or concurrent) validity of TS GOLD through its association with well-established direct assessments of the same or similar learning constructs. The findings of both studies were similar in terms of convergent validity estimates expressed as correlations between the TS GOLD domains and corresponding subscales on norm-referenced assessments, which ranged from low ($r$ = .26 for measures of social-emotional and physical skills in Soderberg et al., 2013) to moderate ($r$ = .68 for measures of literacy and mathematics in Lambert et al., 2015). Additionally, Lambert et al. (2015) performed a confirmatory factor analysis to verify the proposed six domains of TS GOLD. Although the six-factor model produced a statistically significant chi-square value, it also had generally acceptable fit across several indices in a sample of students ages 3 to 5 (standardized root mean square residual [SRMR] = .038, comparative fit index [CFI] = .918, and root mean square error of approximation [RMSEA] = .061), providing support for the internal structure of TS GOLD. Correlations between the latent constructs were positive and strong and ranged from .68 to .93. Finally, Lambert and colleagues (2015) compared the intraclass correlations (ICCs) associated with TS GOLD and external direct assessment, and found the former to be much larger, suggesting "a high probability of rater effects such as leniency or strictness" (p. 60). One limitation of these studies is that they did not report or discuss estimates of discriminant validity. Understanding an assessment's capacity for discrimination is important if these tools are to be used to individualize and guide classroom instruction, as a measure needs to be able

TABLE 1
*Descriptive Statistics of Student Sample and Subsample*

| | Validity subsample | | | Total sample | | |
|---|---|---|---|---|---|---|
| Variable | *N* | *M* | *SD* | *N* | *M* | *SD* |
| Age (in years) | 522 | 5.38 | 0.34 | 1,086 | 5.39 | 0.35 |
| Female | 522 | 0.50 | 0.50 | 1,086 | 0.53 | 0.50 |
| Free- or reduced-price lunch (school level, %)[a] | 502 | 39.79 | 17.65 | 1,020 | 41.36 | 18.15 |
| Individualized education plan (%) | 522 | 0.07 | 0.25 | 1,086 | 0.08 | 0.27 |
| Race-ethnicity[b] | | | | | | |
| White | 522 | 0.73 | 0.44 | 1,086 | 0.73 | 0.45 |
| Black | 522 | 0.16 | 0.36 | 1,086 | 0.16 | 0.37 |
| Hispanic | 522 | 0.05 | 0.22 | 1,086 | 0.05 | 0.22 |
| Asian | 522 | 0.02 | 0.14 | 1,086 | 0.02 | 0.13 |
| Multiracial | 522 | 0.04 | 0.20 | 1,086 | 0.04 | 0.21 |
| Other | 522 | 0.00 | 0.06 | 1,086 | 0.01 | 0.08 |

a. *T* test significant at $p < .05$.
b. Percentages may not total to 100 due to unreported race and overlap between Hispanic ethnicity and racial categories.

to differentiate students' skill levels across learning domains and to differentiate students within a classroom.

### *Purpose of Study*

This study focused on TS GOLD due to its similarities to other performance-based KEAs, its widespread use as a mandated KEA in at least 10 states across the country (see Colorado DOE, 2016; Korobkin, 2012; Louisiana DOE, n.d.; Massachusetts Executive Office of Education, 2016; Michigan DOE, 2016; Minnesota DOE, n.d.; Nevada Early Childhood Advisory Council, 2016; New Jersey DOE, n.d.; State of Alabama DOE, 2015; State of Washington, 2015), and the mixed validity evidence in prior research. We examined TS GOLD for its internal consistency and its convergent and discriminant validity in a sample of kindergarten classrooms. Specifically, the study was designed to answer three primary research questions: Do the previously reported TS GOLD constructs replicate in the current sample? How closely do the TS GOLD ratings of children's skills in key early learning domains at kindergarten entry align with the results of other validated direct assessments of the same or aligned constructs? How well does TS GOLD differentiate children's readiness skills within a classroom? In order to address these questions, we first performed a confirmatory factor analysis and estimated the interdomain correlations for TS GOLD to see how similar/dissimilar students' scores were across the different learning constructs. Second, we examined convergent and divergent validity by comparing children's school readiness scores as measured by teachers using TS GOLD to scores as measured by independent data collectors using direct assessments. Finally, we compared ICCs among student scores as determined by TS GOLD versus the independently administered direct assessments.

TABLE 2
*Descriptive Statistics for Participating Teachers*

| Variable | *N* | *M* | *SD* |
|---|---|---|---|
| Age | 58 | 41.10 | 10.88 |
| Male | 57 | 0.02 | 0.14 |
| Years of teaching experience | 57 | 14.51 | 9.41 |
| Years of education | 58 | 17.19 | 0.99 |
| Bachelor's degree | 58 | 0.43 | 0.50 |
| Master's degree | 58 | 0.57 | 0.50 |
| Ethnicity, White | 58 | 0.91 | 0.30 |
| African American | 58 | 0.05 | 0.24 |
| Multiracial | 58 | 0.03 | 0.19 |

## Research Methods

### *Participants*

Fifty-eight teachers and 1,086 kindergarten students within 16 public elementary schools across four diverse districts in one southeastern state within the United States were included in the present study. These participants were part of a program evaluating best practices of kindergarten readiness assessment. The sample was chosen to contain teachers and children with diverse demographic characteristics (see Tables 1 and 2).

In order to construct the subsample used in the validity analyses, classrooms from the program were strategically selected to maintain diversity, and an average of 10 students within each of these classrooms were randomly selected to receive a battery of previously validated direct assessments. This resulted in a subsample of 522 students in 52 classrooms. Comparisons of the validity subsample to those students not randomly chosen from the larger sample revealed

no significant differences across age, gender, or race ($p > .05$), but the students who received direct assessments did come from schools with a somewhat lower proportion of students receiving free- or reduced-price lunch on average ($p < .05$).

The demographic characteristics of the students in the full sample and validity subsample are presented in Table 1. Participating students were between 4 and 7 years of age upon entering kindergarten, and the majority of children (73%) in the study sample were White. Black students composed the second largest racial group (16 %), Hispanic students represented 10% of the sample, and Asian students accounted for 4% of the sample. Approximately 7% of participants were identified as having special education needs. On average, children included in the sample were situated in schools in which 40% of students qualified for free or reduce-priced lunch.

The background characteristics of all participating teachers are included in Table 2. On average, teachers had been practicing for 14 years. All teachers had earned a bachelor's degree or higher, with more than half of teachers (57%) possessing a master's degree. The vast majority of sampled teachers were White (91%), and almost all were female (98%).

### *Measures*

*TS GOLD.* TS GOLD is a multidimensional early childhood assessment in which teachers observe children's skills during typical instruction for certain period of time, provide documentation of what is observed, and then rate children on a set of items (Heroman et al., 2010). The assessment covers children's development in nine broad skill areas, including literacy, mathematics, language, social-emotional, cognitive, physical, science and technology, social studies, and arts. These skill domains are further described by 38 objectives (listed by skill domain in Table 3) and 65 behavioral indicators referred to as "dimensions." Using an online portal, teachers collect pieces of evidence of students' demonstrated skills, including classroom notes, videos, and samples of student work related to multiple domains of school readiness. At several points during the year, teachers are asked to review their students' documentation and rate them using "age-bands" to determine whether children's skills are developing in expected ways. In the first six skill areas listed (literacy through physical), teachers rate children's skill levels on each of the dimensions using a 9-point scale with descriptive anchors at points 2, 4, 6, and 8 on the scale. These ratings are placed within a continuum that reflects designated developmental expectations within each learning objective based on nationally normed benchmarks (Teaching Strategies, 2012). Previous studies of TS GOLD have shown moderate to strong reliability (Lambert et al., 2015) and adequate to strong convergent validity (Lambert et al., 2015; Soderberg et al., 2013).

TABLE 3
*Teaching Strategies GOLD Domains and Objectives*

| Domain | Objectives |
|---|---|
| Literacy | Demonstrates phonological awareness |
| | Demonstrates knowledge of the alphabet |
| | Demonstrates knowledge of print and its uses |
| | Comprehends and responds to books and other texts |
| | Demonstrates emergent writing skills |
| Language | Listens to and understands increasingly complex language |
| | Uses language to express thoughts and needs |
| | Uses appropriate conversational and other communication skills |
| Math | Uses number concepts and operations |
| | Explores and describes spatial relationships and shapes |
| | Compares and measures |
| | Demonstrates knowledge of patterns |
| Cognitive | Demonstrates positive approaches to learning |
| | Remembers and connects experiences |
| | Uses classification skills |
| | Uses symbols and images to represent something not present |
| Social-emotional | Regulates own emotions and behaviors |
| | Establishes and sustains positive relationships |
| | Participates cooperatively and constructively in group situations |
| Physical | Demonstrates traveling skills |
| | Demonstrates balancing skills |
| | Demonstrates gross-motor manipulative skills |
| | Demonstrates fine-motor strength and coordination |

*Note*. The six domains above represent the subset of the nine domains covered by Teaching Strategies GOLD. These were included in the present study.

*Direct assessments of literacy, mathematics, and language*. Literacy, mathematics, and expressive language were assessed using subtests from the Woodcock-Johnson III Tests of Achievement (WJTA; Woodcock, McGrew, & Mather, 2001). All direct assessments and their corresponding TS GOLD domain are listed in Table 4. Descriptive statistics for all measures in the current sample are provided in Table 5. The WJTA is a widely used (e.g., Burchinal, Peisner-Feinberg, Pianta, & Howes, 2002; Duncan et al., 2007; Peisner-Feinberg et al., 2001), individually administered assessment battery that measures achievement in individuals from age 2 through adulthood. Twelve achievement subtests can be used with young children, with items (scored as either incorrect or

TABLE 4
*Domains and Skills in Direct Assessments*

| Domain | Measure | Skills assessed |
|---|---|---|
| Literacy[a] | WJ Letter-Word Identification | Identification and recognition of letters |
| | | Identification of printed words |
| | WJ Word Attack | Pronunciation of phonically regular nonwords (phonological awareness) |
| Language | WJ Picture Vocabulary | Identification of objects |
| Math[b] | WJ Applied Problems | Analyze and solve math problems |
| | WJ Quantitative Concepts | Identification of math terms |
| | | Identification of number patterns |
| Cognitive | Pencil Tap | Inhibitory control |
| | Head-Toes-Knees-Shoulders | Remember increasingly complex instructions |
| | | Inhibitory control |
| | | Ability to focus on examiner's instructions |

*Note*. The six domains above represent the subset of the nine domains covered by Teaching Strategies GOLD included in the present study. WJ = Woodcock-Johnson.
a. The two subtests used to assess literacy compose the WJ Tests of Achievement Basic Reading Composite.
b. The two subtests used to assess mathematics compose the WJ Tests of Achievement Math Reasoning cluster.

TABLE 5
*Descriptive Statistics of Measures*

| Measure | *N* | *M* | *SD* | Score range |
|---|---|---|---|---|
| TS GOLD Literacy | 522 | 60.66 | 7.80 | 7.00–97.00 |
| TS GOLD Language | 522 | 49.31 | 14.34 | 20.00–70.00 |
| TS GOLD Math | 522 | 37.09 | 7.84 | 0.00–57.00 |
| TS GOLD Cognitive | 522 | 54.56 | 10.05 | 16.00–86.00 |
| TS GOLD Social-emotional | 522 | 53.17 | 9.57 | 9.00–79.00 |
| TS GOLD Physical | 521 | 33.06 | 4.64 | 12.00–42.00 |
| WJ Basic Reading[a] | 521 | 392.13[b] | 26.52 | 328.00–507.00 |
| WJ Picture Vocabulary | 520 | 473.66 | 9.35 | 440.00–505.00 |
| WJ Math Reasoning[c] | 516 | 432.80 | 14.58 | 381.00–473.00 |
| Pencil Tap | 522 | 13.92 | 3.41 | 0.00–16.00 |
| HTKS | 522 | 26.08 | 12.54 | 0.00–40.00 |

*Note*. TS = Teaching Strategies; WJ = Woodcock-Johnson; HTKS = Head-Toes-Knees-Shoulders.
a. Basic Reading is a composite of Letter-Word Identification and Word Attack subscales.
b. All WJ scores are reported as W scores.
c. Math Reasoning is a composite of Applied Problems and Quantitative Concepts subscales.

correct) progressing in difficulty over the course of the assessment. Six of the WJTA subtests were used in this study due to their alignment with the constructs covered by TS GOLD. Picture Vocabulary was used to assess students' language skills, in which children are presented with an illustration and then asked to point to a particular object or state what is depicted on the card. Literacy was assessed using the Letter-Word Identification and Word Attack subtests. Letter-Word Identification is a measure of children's sight vocabulary. This subtest begins with letters and moves to increasingly complex words, measuring children's ability to identify words either through decoding or visual memory. Word Attack is a measure of the ability to decode and pronounce phonically regular nonwords, or pseudo words. These two subtests combine to form a Basic Reading composite score used for analyses. Mathematics skills were assessed with the Applied Problems and Quantitative Concepts subtests. Applied Problems measures analytical and problem-solving skills and requires children to listen to a problem, identify the procedure to solve it, and perform simple calculations. Quantitative Concepts measures math knowledge of concepts, symbols, and vocabulary and includes two sections: Concepts (recognizing numbers, shapes, and sequences) and Number Series (identifying the missing digit in a series). Together, Applied Problems and Quantitative Concepts compose the Math Reasoning cluster used for analyses.

Previous research on the WJTA has established its validity and reliability in diverse, nationally representative samples of children and adolescents (Woodcock et al., 2001; Woodcock, McGrew, Schrank, & Mather, 2007). A series of analyses provide information about each subtest's psychometric properties. Median reliability coefficients clustered by intended use of assessment and age of test taker ranged from .85 to .99, with the majority of subtests possessing reliabilities above .90 (Woodcock et al., 2001, 2007). Several studies in samples of different ages and abilities also provided strong evidence for the concurrent, discriminant, and predictive validity (McGrew & Woodcock, 2001).

*Direct assessments of self-regulation.* Aspects of children's behavioral regulation/executive functioning (e.g., the integration of children's behavior requiring attention to instructions, working memory, and inhibitory control; Cameron,

McClelland, Matthews, & Morrison, 2009) were assessed using two measures: Pencil Tap (Diamond & Taylor, 1996; Smith-Donald, Raver, Hayes, & Richardson, 2007) and the Head-Toes-Knees-Shoulders (HTKS) task (Cameron et al., 2009). Both assessments have been widely used in recent developmental and early education research.

Pencil Tap was adapted from the Preschool Self-Regulation Assessment (PSRA; Smith-Donald et al., 2007) to measure inhibitory control, or a child's ability to stop himself or herself from the dominant response. In Pencil Tap, children are given instructions to tap their pencil once on the table when the examiner taps twice, and twice when the assessor taps once. In the present study, raw scores were calculated by summing all correct and incorrect responses, with possible totals ranging from 0 to 16. Previous research has provided evidence of strong reliability (Smith-Donald et al., 2007) and concurrent and predictive validity (Blair & Razza, 2007; Rimm-Kaufman, Curby, Grimm, Nathanson, & Brock, 2009; Smith-Donald et al., 2007) of Pencil Tap in samples of young children.

The HTKS task is a measure of children's behavioral regulation and assesses a child's ability to control his or her behavior through the use of attention, working memory, inhibitory control, and cognitive flexibility (Cameron et al., 2009). Framed as game, the task asks children to do one of two actions that is the opposite of what the assessor says (e.g., when examiner says, "Touch your head," the student should touch his or her toes; when examiner says, "Touch your toes," child should touch his or her head). As the examinee progresses through the three levels of the assessment, the instructions become increasingly complex with the addition of new rules. Each item is scored as an incorrect response (0), a self-correct (1), or a correct response (2), for a maximum score of 40. Raw scores were used for all analyses. The HTKS task has demonstrated strong reliability (Cameron et al., 2008) and construct (Cameron et al., 2008, 2009) and predictive validity (McClelland et al., 2007) in other studies.

### *Procedures*

*TS GOLD.* Before using TS GOLD in their classrooms, all teachers participated in an onsite 2-day training provided by Teaching Strategies. The session focused on various aspects of administration, including navigation of the online application, documentation of student work, and assessment of children's readiness skills using the TS GOLD criteria. Following training, teachers were instructed to observe and document children's readiness skills for the first 4 weeks of school and to subsequently assess children's skills during a 2-week assessment window. Teachers reported a wide range of administration times via a post-assessment survey. For the majority of participants, the TS GOLD documentation took an average of 1 to 2 hours per child. Teachers also reported that entering scores for the fall assessment took an additional 30 minutes to 1 hour per student.

*Direct assessments.* Eleven independent data collectors completed the direct assessments on-site using an electronic system. All data collectors completed a 2-day training led by the research team. In order to ensure fidelity to assessment procedures, data collectors were required to practice administration and record one session for review and feedback by a university researcher. Data collectors were also supervised in their initial administration. Direct assessments were conducted within 6 to 10 weeks from the start of student instruction and overlapped with the TS GOLD assessment window. The full assessment battery took between 20 and 45 minutes per student to complete.

### *Analytic Approach*

We performed several different analyses in order to address our three primary research questions. First, we verified the proposed six-factor structure of TS GOLD in a confirmatory factor analysis (CFA) with the full sample using Mplus Version 7.11. In a CFA, the unobserved, theoretical constructs are modeled to predict the observed responses to the assessment items to which they are hypothetically related. Multiple fit indices are considered to determine if the proposed model provides an acceptable fit to the observed data, with adequate fit on multiple (but not all) indicators suggesting an appropriate modeling of the underlying data structure. For the present study, the CFA model accounted for the clustering of students in classrooms (using the "type=complex" command in Mplus) and included six factors representing the primary learning domains suggested by TS GOLD: social-emotional, physical, language, cognitive, literacy, and math. Each item in the assessment was allowed to load onto its corresponding domain but had no links to any of the other skill areas. The domains were allowed to covary, but no specific inter-item covariances were included. We examined several indices in order to determine the adequacy of the fit using Hu and Bentler's (1999) criteria for acceptable fit: chi-square test ($p > .05$), RMSEA (<.06), SRMR (<.08), CFI (>.95), and Tucker Lewis index (TLI; >.95).

To further examine the discriminant validity of the proposed factors, we examined the interdomain correlations produced by a two-level CFA (using the "type=twolevel" option in Mplus) and calculated the average variance extracted (AVE) for each latent construct (Fornell, Tellis, & Zinkhan, 1982). The two-level factor analysis, in which we estimated separate models at the within- and between-class levels, permitted us to determine the extent to which observed correlations between latent constructs were reflective of scoring similarities within a classroom (e.g., within cluster) and/or across teachers (e.g., between cluster). We used the aforementioned indices ($\chi^2$, RMSEA, SRMR, CFI, TLI) to establish the adequacy of the fit for two-level factor model. The AVE represents the average amount of variance

explained by a given construct ($\xi_j$) and is calculated by taking the average of all the squared standardized factor loadings ($\lambda^2$) across *k* items that load onto the construct:

$$\text{AVE}\xi_j = \frac{1}{K_j}\sum_{k=1}^{K_j}\lambda_{jk}^2.$$

Under the criterion suggested by Fornell et al. (1982), a construct has exhibited discriminant validity if the square root of the AVE for that construct is larger than the absolute value of the correlation between that construct (*j*) and any of the other latent constructs (*i*) included in the model:

$$\sqrt{\text{AVE}\xi_j} > max\left|r_{ij}\right|.$$

More simply, this formula ascertains whether or not the construct is more strongly related to the items that compose it than to other constructs.

In the remaining analyses, we restricted our sample to include only those students who were assessed using both TS GOLD and the battery of direct assessments. In order to evaluate the convergent and discriminant validity as well as differentiation among TS GOLD domains, we estimated a series of bivariate correlation coefficients using Mplus Version 7.11. All correlations accounted for the nesting of students within classrooms. To assess convergent validity, we examined the correlation coefficients between TS GOLD and direct assessments of the same skill area. Typically, correlations between performance-based ratings and direct assessments of similar constructs are moderate, ranging from .40 to .60 (Cabell et al., 2009; Kilday et al., 2012; Sudkamp et al., 2012). To further explore TS GOLD's ability to distinguish between skill levels across the different learning domains, we looked at the associations between the TS GOLD domains and direct assessments of different learning constructs (e.g. TS GOLD literacy score vs. direct assessment math score). Although there is not currently a standard for the desired magnitude of these correlations, the relationship between two assessments that target different learning constructs should be weaker than the relationship exhibited between assessments of the same or similar domains (D. Campbell & Fiske, 1959; Downing, 2003; Messick, 1995; Peter, 1981).

ICCs provided additional information about TS GOLD's ability to differentiate among students within a given learning domain in the same classroom. Similar to the two-level CFA, the ICC partitions the amount of variance in students' scores that is between the clusters (in this case, the classroom/teacher) versus within the clusters. It is defined by the following ratio:

$$\frac{\sigma_B^2}{\sigma_B^2+\sigma_W^2},$$

where $\sigma_B^2$ represents the between-cluster variance and $\sigma_W^2$ represents the within-cluster variance. In other words, the ICC compares that amount of score variation that can be attributed to the classroom/teacher to the total amount of variance that is observed in scores. By definition, ICCs range from 0 to 1, with higher values indicating that students' scores look more similar within the classroom than across classrooms. Although there is no convention by which to judge the magnitude of the ICC, previous research on ICCs in performance-based assessment found values that ranged from 0.15 to 0.35 (Mashburn et al., 2006; Waterman et al., 2012). The ICC coefficients were calculated using a two-step process in Stata 14. First, a series of two-level mixed-effects models were estimated using maximum likelihood estimation, with scores for each TS GOLD domain and direct assessment as the dependent variables and students nested within classrooms. Through this estimation, we were able to obtain the two components needed to calculate the ICC: the residual variance ($\sigma_W^2$) and the variance associated with the classroom ($\sigma_B^2$). The ICC was then calculated using the formula above.

## Results

### *Internal Structure of TS GOLD*

We first confirmed the six-factor structure of TS GOLD using CFA. On the basis of the criteria suggested by Hu and Bentler (1999), the estimated six-factor model had acceptable fit on two (RMSEA = .043, SRMR = .052) of the five indices, $\chi^2(1209) = 3586$, CFI = .845, TLI = .837. All item loadings were significant, and standardized factor loadings ranged from .54 to .89. The correlation coefficients between each of the latent factors (Table 6) were positive and ranged from moderate ($r$ = .66 between physical and literacy) to large ($r$ = .91 between language and cognitive) based on the criteria suggested by Hinkle, Wiersma, and Jurs (2003). Although all of the associations among the direct assessments (Table 7) were also significant and positive, the coefficients were smaller in magnitude, ranging from small ($r$ = .17 for literacy with self-regulation) to moderate ($r$ = .61 for literacy with math).

Because of the nested nature of our TS GOLD data, it is important to determine the extent to which the aforementioned correlations among TS GOLD domains were reflective of scoring relationships within or between classes. In order to do this, we examined the correlations of the latent constructs produced in a two-level CFA. The two-level model exhibited acceptable fit on two (RMSEA = .039, SRMR = .043 at within level) of the five indices, $\chi^2(2418) = 6287$, CFI = .906, TLI = .901. All items loaded significantly onto their given constructs ($p < .05$), with the exception of one item in math and one in literacy at the between level ($p < .10$). Standardized factor loadings ranged from .59 to .90 for the within-level and from .28 to .99 for the

TABLE 6
*Correlations Between Latent Factors (off diagonal) and Square Root of the Average Variance Extracted (diagonal)*

| Domain | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1. Social-emotional | *.812* | | | | | |
| 2. Physical | .769 | *.789* | | | | |
| 3. Language | .854 | .794 | *.827* | | | |
| 4. Cognitive | .856 | .757 | .908 | *.834* | | |
| 5. Literacy | .666 | .662 | .812 | .833 | *.768* | |
| 6. Mathematics | .691 | .676 | .826 | .847 | .839 | *.772* |

*Note*. All of the correlations are significant ($p < .001$). The coefficients along the diagonal represent the square root of average variance extracted, which is equivalent to the average of the squared factor loadings for each construct.

TABLE 7
*Correlations of Direct Assessments*

| Assessment | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1. WJ Basic Reading | — | | | | |
| 2. WJ Picture Vocabulary | .475 | — | | | |
| 3. WJ Math Reasoning | .614 | .511 | — | | |
| 4. Pencil Tap | .173 | .230 | .368 | — | |
| 5. Heads-Toes-Knees-Shoulders | .300 | .404 | .534 | .404 | — |

*Note*. All correlations are significant ($p < .001$). WJ = Woodcock-Johnson.

between-level model. Both the within- and between-cluster interdomain correlations were moderate to large, ranging from .67 (math with social-emotional) to .94 (math with literacy) for the within-class correlations and from .69 (math with literacy) to .94 (social-emotional with language) for the between-class correlations. The relatively consistent magnitude of the correlations at both levels suggests that, on average, children's ratings within a given classroom and a teacher's average classroom score tended to look similar across all TS GOLD domains.

The relationships between TS GOLD domains were further explored using the square root of the AVE for each latent construct (included with factor correlations in Table 6). All latent constructs had at least one estimated correlation with another domain that exceeded the square root of the AVE across all of the proposed TS GOLD domains. In particular, the square roots of the AVE for literacy cognitive, and mathematics domains are smaller than three of the five correlations with the other latent constructs. This suggests that more of the observed variation in the literacy items is explained by the language, cognitive, or mathematics latent constructs than by the literacy construct.

### *Convergent and Discriminant Validity With Direct Assessments*

To establish the convergent validity of TS GOLD, we examined the relationship between each domain and the direct assessment of the same construct, represented by the bolded correlation coefficients in Table 8. The strength of the within-domain associations varied based on the skill area and ranged from small to moderate. The TS GOLD literacy domain demonstrated the strongest correlation with its corresponding direct assessment ($r = .68$), whereas the cognitive domain displayed the weakest associations with the direct assessments of self-regulation ($r = .28$ and $r = .33$ for Pencil Tap and HTKS, respectively).

TABLE 8
*Correlations Between TS GOLD Domains and Direct Assessments*

| | TS GOLD domains | | | |
|---|---|---|---|---|
| Direct assessment | Literacy | Language | Math | Cognitive |
| WJ Basic Reading | **.675** | .442 | .436 | .388 |
| WJ Picture Vocabulary | .472 | **.400** | .336 | .339 |
| WJ Math Reasoning | .671 | .517 | **.558** | .494 |
| Pencil Tap | .283 | .293 | .235 | **.284** |
| Heads-Toes-Knees-Shoulders | .399 | .337 | .312 | **.326** |

*Note*. All correlations are significant ($p < .001$). TS = Teaching Strategies; WJ = Woodcock-Johnson. Bolded correlations indicate corresponding assessments and domains. For instance, WJ Basic Reading is most closely aligned with the TS GOLD domain of literacy.

Next, we examined the associations with TS GOLD domain scores and direct assessments of *different* learning domains for evidence of discriminant validity. We found that the majority of the correlations across less similar constructs were comparable in strength to the relationships within the same skill area. For instance, the TS GOLD literacy domain had a moderate relationship of similar magnitude to math compared to its within-domain association with Woodcock-Johnson Basic Reading ($r = .67$ vs. $r = .68$ for math and reading, respectively). In the case of both TS GOLD language and cognitive domains, several of the cross-domain assessments demonstrated stronger correlations than the corresponding measure of the same construct. These findings also suggest that many of the TS GOLD domain scores had as much or more in common with measures of different skills than they did with assessments of the same learning domain.

### *Skill-Level Differentiation Within TS GOLD Domains*

To determine the extent to which teachers' ratings discriminated among children's skill levels in a given learning domain relative to the direct assessments, we examined the ICC coefficients. The ICCs for each TS GOLD domain and direct assessment are presented in Table 9. The values derived from the TS GOLD domain-specific scores ranged from .18 in literacy to .59 in the physical domain, with a median ICC of .37. Thus, on average, 37% of a student's TS GOLD score within a learning domain can be explained by

TABLE 9
*Intraclass Correlation (ICC) Coefficients*

| Domain | TS GOLD | Direct assessment |
|---|---|---|
| Literacy | .185 | .017 |
| Language | .277 | .019 |
| Math | .422 | .039 |
| Cognitive | .369 | .014[a] |
| | — | .036[b] |
| Social-emotional | .369 | — |
| Physical | .591 | — |

*Note*. ICCs were estimated as the ratio of between-class variance to total score variance. Not all TS GOLD domains have corresponding direct assessments and the Cognitive domain corresponds to two direct assessments. TS = Teaching Strategies.
a. ICC for Pencil Tap.
b. ICC for Head-Toes-Knees-Shoulders.

the student being in a particular classroom. In comparison, the ICCs corresponding to the direct assessments were much smaller (.01 for Pencil Tap to .04 for math and HTKS), indicating that the variability in students' direct assessment scores was almost entirely attributable to differences between students, with 1% to 4% of the variation explained by a student being in a particular classroom.

## Discussion

The large-scale use of KEAs has evolved over the past decade to include more states, domains of readiness, and forms of utilization. Most commonly, states have relied on embedded, performance-based assessments to provide information about students' skills at school entry. This type of assessment has appealed to educational administrators and practitioners for multiple reasons, including (a) the increased validity of the results through the observation of students' exhibited skills in genuine contexts while still maintaining adequate psychometric properties, (b) the enhanced ability of such assessments to comprehensively inform instruction and intervention, and (c) the flexibility of administration without sacrificing instructional time. However, such performance-based assessments have also been criticized as being more susceptible to the influence of rater characteristics (Jonsson & Svingby, 2007), meaning that variation in students' scores may be driven by differences across administrators rather than actual differences in students' abilities (Engelhard, 1994; Waterman et al., 2012). In our examination of one such assessment, TS GOLD, we found important strengths and weaknesses that both align with the literature and diverge from the purported advantages of embedded, performance-based assessment.

Consistent with previous research (Burts & Kim, 2014; Cabell et al., 2009; Lambert et al., 2014; Soderberg et al., 2013), TS GOLD exhibited a consistent internal structure composed of the proposed readiness domains. It also demonstrated strong associations with independent direct assessments of the same constructs. Thus, the present study provides further evidence of the scale reliability in terms of factor structure as well as the convergent validity.

However, the strength of TS GOLD's psychometric characteristics diminished somewhat when we considered both discriminant validity and the potential for rater effects, which have not been examined in prior research. For a measure to have evidence of discriminant validity, we would expect associations between conceptually related domains to be stronger than the relationships between measures of different constructs (D. Campbell & Fiske, 1959; Downing, 2003; Messick, 1995; Peter, 1981). For example, we would anticipate that two measures of literacy would have a larger correlation than a measure of literacy and a measure of mathematical ability, as the former result would suggest that the measure is assessing skills that are uniquely associated with literacy, whereas the latter implies the influence of a confounding factor also captured by the theoretically unrelated math assessment. Although several of the TS GOLD domains were strongly related to the direct assessments of corresponding skills, three of the four domains (literacy, language, and cognitive) had comparable or larger associations with direct assessments of different constructs (see Table 8). The lack of discrimination was further indicated by the high correlations between the TS GOLD learning domains in our factor analysis relative to the average variance extracted by the individual constructs. This suggests that teachers tended to rate individual students more similarly across all learning constructs despite empirical evidence of more substantial variation across domains when skills are measured via direct assessment. Additionally, TS GOLD appeared less able to differentiate students' skills in a specific learning area within a classroom. The substantially larger ICCs associated with TS GOLD ratings indicated that students' scores in each learning construct were much more similar within a given classroom compared to those produced by the direct assessments. This suggests that the readiness ratings produced using TS GOLD—and likely other performance-based assessments—may be more influenced by factors that are separate from a child's actual skills compared to results obtained from direct assessments (Mashburn et al., 2006; Waterman et al., 2012).

Measuring the readiness competencies of incoming kindergarten children is a challenging task. No assessment will satisfy all of the potential uses for entry assessment, and because no assessment is a panacea for all instructional and programmatic needs, educational policymakers must be deliberate about matching appropriate measures to their constituents' identified requirements. Although TS GOLD was originally intended as a curriculum-embedded assessment (Halle et al., 2011), it is now commonly utilized, as well as other performance-based measures, with a variety of curricula as a KEA and as a progress-monitoring tool in preschool (Teaching Strategies, 2013b). It is important for educational policymakers and educators to understand whether the

psychometric properties of the chosen assessments allow for the data to be used in the way the user intends. If teachers are investing additional time in order to assess a wide range of skills for all children in their classrooms with the intention of individualizing instruction, the expectation is that they come away with an improved understanding of how students are differentially functioning across multiple learning domains. Thus, the value of having an assessment that measures multiple types of skills is reduced if children tend to look the same across readiness domains.

### Limitations and Directions for Future Research

There are several important limitations that affect the generalizability of the current findings. First, the performance of a measure is influenced by context (Halle et al., 2011). As emphasized above, this study considers only the use of TS GOLD as a KEA, with the intention of identifying students who are at risk and in need of further assessment and/or early intervention in a specific area. As described in the introduction, TS GOLD and other performance-based measures are now being used for this purpose. We did not examine the use of TS GOLD as a formative assessment linked to implementation of a particular curriculum as this was not the purpose of this study.

Additionally, the broad set of behaviors assessed within some of the TS GOLD domains made it difficult to identify the appropriately aligned direct assessments. A lack of alignment between the constructs assessed could lead to lower convergent-validity associations. This was particularly problematic for the cognitive and social-emotional domains, the first of which touches on a wide range of skills, from approaches to learning and working memory to symbolic representation and classification skills, and overlaps with the behaviors represented within social-emotional domain. We chose to look at the relationship between cognitive scores and those obtained from HTKS and Pencil Tap as opposed to the social-emotional domain due to the narrower focus of the cognitive domain as well as better alignment with the behaviors measured by the direct assessments (in particular, behavioral regulation and executive functioning). Still, this potential lack of alignment was less problematic for the academic domains, and the correlations between the direct assessments and TS GOLD in these domains were largely of the size and significance needed to establish convergent validity.

Furthermore, the data were collected during the pilot year of TS GOLD after a single 2-day, in-person training session, which could suggest that participating teachers had limited familiarity with the measure. This lack of experience and comfort with TS GOLD and its procedures is underscored by the small number of teachers attempting and/or achieving certified reliability within the present sample. Research suggests that the accuracy of teachers' ratings of students could improve with continued use of an instrument (Ackerman & Coley, 2012; Meisels, Wen, & Beachy-Quick, 2010), although other studies have found that familiarity only moderately improves the accuracy of teacher-administered performance-based assessments (Begeny & Buchanan, 2010). Further research should be done in classrooms where teachers have implemented the measure of interest for multiple years to see if the similar patterns hold and to the same degree found within this sample.

Finally, additional research is needed to understand the utility of performance-based assessments and the most appropriate uses of these types of measures. In the present study, we found that teachers scored students' skills much more similarly across school readiness domains using TS GOLD than was evident when students were measured using direct assessments. For example, when teachers rated children in their classroom highly in literacy skills, they also tended to rate them highly in math skills. Although the level of young children's skills across readiness domains tends to be somewhat consistent within a child, we would expect less consistency in children's skills across readiness domains than was demonstrated when children were assessed by teachers using TS GOLD. However, it is unclear from the current analyses whether or not teachers systematically and consistently differ in the ways in which the rate their students. Future studies should do more to parse out rater biases based on administrator and child characteristics.

### Acknowledgments

This study was supported by a grant from Elevate Early Education (E3). The findings, opinions, and implications expressed are those of the authors and not E3. The authors are grateful for the help of the many children, teachers, administrators, and field staff who participated in this study. We also thank Elise Rubinstein, Jason Downer, Bridget Hamre, Robert Pianta, Caroline Werenskjold, and Lee Williams for their contributions to this project.

### References

Ackerman, D. J., & Coley, R. J. (2012). *State pre-K assessment policies: Issues and status*. Policy information report. Princeton, NJ: Educational Testing Service.

Alaska Department of Education and Early Development. (2016). *Alaska Development Profile implementation guide*. Retrieved from https://education.alaska.gov/tls/Assessments/DevelopmentalProfile/Fall2016/ImplementationGuide_July2016.pdf

Arkansas Department of Education. (n.d.) *K-2 assessment*. Retrieved from http://www.arkansased.gov/divisions/learning-services/assessment/k-2-assessment

Atkins-Burnett, S. (2007). *Measuring children's progress from preschool through third grade* (No. 2d7310d35a3e4a129792d-de6d1f5107b). Washington, DC: Mathematica Policy Research. Retrieved from: https://www.mathematica-mpr.com/-/media/publications/pdfs/measchildprogress.pdf

Bagnato, S. J., & Macy, M. (2010). Authentic assessment in action: A "REAL" solution. *NHSA DIALOG*, *13*(1), 42–45.

Begeny, J. C., & Buchanan, H. (2010). Teachers' judgments of students' early literacy skills measured by the Early Literacy Skills Assessment: Comparisons of teachers with and without assessment administration experience. *Psychology in the Schools*, *47*(8), 859–868.

Belfield, C. R., Nores, M., Barnett, S., & Schweinhart, L. (2006). The High/Scope Perry Preschool Program cost-benefit analysis using data from the age-40 follow up. *Journal of Human Resources*, *41*, 162–190. http://dx.doi.org/10.3368/jhr.XLI.1.162

Belsky, J., & MacKinnon, C. (1994). Transition to school: Developmental trajectories and school experiences. *Early Education and Development*, *5*, 106–119. doi:10.1207/s15566935eed0502_3

Blair, C., & Raver, C. (2015). School readiness and self-regulation: A developmental psychobiological approach. *Annual Review Psychology*, *66*, 711–731. doi:10.1146/annurev-psych-010814-015221

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.

Brooks-Gunn, J., & Markman, L. B. (2005). The contribution of parenting to ethnic and racial gaps in school readiness. *Future of Children*, *15*, 139–168. Retrieved from http://www.jstor.org/stable/1602666

Burchinal, M. R., Peisner-Feinberg, E., Pianta, R., & Howes, C. (2002). Development of academic skills from preschool through second grade: Family and classroom predictors of developmental trajectories. *Journal of School Psychology*, *40*, 415–436.

Burts, D., & Kim, D. (2014). The Teaching Strategies GOLD® assessment system: Measurement properties and use. *Dialog*, *17*, 122–135.

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, Speech, and Hearing Services in Schools*, *40*, 161–173.

California Department of Education. (2016). *Introduction to Desired Results*. Retrieved from http://www.cde.ca.gov/sp/cd/ci/desiredresults.asp

Cameron, C., McClelland, M. M., Jewkes, A. M., Connor, C. M., Farris, C. L., & Morrison, F. J. (2008). Touch your toes! Developing a direct measure of behavioral regulation in early childhood. *Early Childhood Research Quarterly*, *23*, 141–158.

Cameron, 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.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81.

Campbell, F. A., & Ramey, C. T. (1994). Effects of early intervention on intellectual and academic achievement: A follow-up study of children from low-income families. *Child Development*, *65*(2), 684–698.

Campbell, F. A., & Ramey, C. T. (1995). Cognitive and school outcomes for high-risk African-American students at middle adolescence: Positive effects of early intervention. *American Educational Research Journal*, *32*(4), 743–772.

Cash, A. H., Hamre, B. K., Pianta, R. C., & Myers, S. S. (2012). Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly*, *27*, 529–542. doi:10.1016/j.ecresq.2011.12.006

Colorado Department of Education. (2016). *Memorandum of understanding with Teaching Strategies® regarding the use of GOLD® for kindergarten school readiness*. Retrieved from https://www.cde.state.co.us/schoolreadiness/tsg_cde_mou_12_21_2015

Connecticut Department of Education. (n.d.). *Fall Kindergarten Entrance Inventory*. Retrieved from http://www.csde.state.ct.us/public/csde/cedar/assessment/kindergarten/fall.htm

Connors-Tadros, L. (2014). *Information and resources on developing state policy on kindergarten entry assessment (KEA) (CEELO FASTFacts)*. New Brunswick, NJ: Center on Enhancing Early Learning Outcomes. Retrieved from http://ceelo.org/wp-content/uploads/2014/02/KEA_Fast_Fact_Feb_11_2014_2.pdf

Conroy, M., & Brown, W. (2004). Early identification, prevention, and early intervention with young children at risk for emotional or behavioral disorders: Issues, trends, and a call for action. *Behavioral Disorders*, *29*, 224–236.

Dennis, L. R., Rueter, J. A., & Simpson, C. G. (2013). Authentic assessment: Establishing a clear foundation for instructional practices. *Preventing School Failure: Alternative Education for Children and Youth*, *57*(4), 189–195. doi:10.1080/1045988X.2012.681715

Diamond, A., & Taylor, C. (1996). Development of an aspect of executive control: Development of the abilities to remember what I said and to "Do as I say, not as I do." *Developmental Psychobiology*, *29*(4), 315–334.

Downing, S. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37*, 830–837.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*, 1428–1446.

Dunlap, G., Johnson, L. F., & Robbins, F. R. (1990). Preventing serious behavior problems through skill development and early intervention. In A. C. Repp & N. N. Singh (Eds.), *Perspectives on the use of nonaversive and aversive interventions for persons with developmental disabilities* (pp. 273–286). Sycamore, IL: Sycamore Publishing Company.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93–112. doi:10.1111/j.1745-3984.1994.tb00436.x

Ferrer, E., Shaywitz, B. A., Holahan, J. M., Marchione, K. E., Michaels, R., & Shaywitz, S. E. (2015). Achievement gap in reading is present as early as first grade and persists through adolescence. *Journal of Pediatrics*, *167*, 1121–1125.

Florida Office of Early Learning. (2014). *Florida Kindergarten Readiness Screener (FLKRS) and Voluntary Prekindergarten (VPK) education program accountability: An overview*. Retrieved from http://www.floridaearlylearning.com/sites/www/Uploads/files/Providers/VPK/FLKRS_VPK_Accountability_Overview_11-07-14.pdf

Fornell, C., Tellis, G. J., & Zinkhan, G. M. (1982). Validity assessment: A structural equations approach using partial

least squares. In B. Walker et al. (Eds.), *An assessment of marketing thought and practice* (pp. 405–409). Chicago, IL: AMA.

Galindo, C., & Sonnenschein, S. (2015). Decreasing the SES math achievement gap: Initial math proficiency and home learning environments. *Contemporary Educational Psychology*, *43*, 25–38. doi:10.1016/j.cedpsych.2015.08.003

Halle, T., Zaslow, M., Wessel, J., Moodie, S., & Darling-Churchill, K. (2011). *Understanding and choosing assessments and developmental screeners for young children: Profiles of selected measures*. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Hamre, B. K., & Pianta, R. C. (2001). Early teacher–child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development*, *72*(2), 625–638.

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

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston, MA: Houghton Mifflin.

Hoyt, W. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*, *5*, 64–86. doi:10.1037//1082-989X.5.1.64

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Illinois State Board of Education. (2016). *Early childhood education: Kindergarten corner*. Retrieved from http://www.isbe.state.il.us/earlychi/html/kindergarten_corner.htm?col2=open#CollapsiblePanel2

Isaacs, J. B. (2012). *Starting school at a disadvantage: The school readiness of poor children*. Retrieved from http://www.brookings.edu/research/papers/2012/03/19-school-disadvantage-isaacs

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*, 130–144. doi:10.1016/j.edurev.2007.05.002

Karelitz, T. M., Parrish, D. M., Yamada, H., & Wilson, M. (2010). Articulating assessments across childhood: The cross-age validity of the Desired Results Developmental Profile–Revised. *Educational Assessment*, *15*, 1–26. doi:10.1080/10627191003673208

Kilday, C. R., Kinzie, M. B., Mashburn, A. J., & Whittaker, J. V. (2012). Accuracy of teacher judgments of preschoolers' math skills. *Journal of Psychoeducational Assessment*, *30*(2), 148–159.

Kim, D. H., & Smith, J. D. (2010). Evaluation of two observational assessment systems for children's development and learning. *NHSA Dialog*, *13*(4), 253–267.

Koretz, D. M. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources*, *37*(4), 752–777.

Korobkin, M. (2012). *Delaware launches early learning survey*. Wilmington, DE: Rodel Foundation of Delaware. Retrieved from http://www.rodelfoundationde.org/delaware-launches-early-learning-survey/

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.

Lambert, R. G., Kim, D. H., & Burts, D. C. (2015). The measurement properties of the Teaching Strategies GOLD® assessment system. *Early Childhood Research Quarterly*, *33*, 49–63.

Lee, V. E., & Burkam, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school*. Washington, DC: Economic Policy Institute.

Little, M. H., Cohen-Vogel, L., & Curran, F. C. (2016). Facilitating the transition to kindergarten. *AERA Open*, *2*(3). doi:2332858416655766

Louisiana Department of Education. (n.d.). *Birth to kindergarten entry assessment tool Teaching Strategies GOLD®*. Retrieved from http://www.louisianabelieves.com/docs/default-source/early-childhood/training-description-birth-to-kindergarten-entry-assessment-tool.pdf?sfvrsn=4

Macy, M., & Bagnato, S. J. (2010). Keeping it "REAL" with authentic assessment. *NHSA Dialog*, *13*(1), 1–20.

Magnuson, K. A., Meyers, M. K., Ruhm, C. J., & Waldfogel, J. (2004). Inequality in preschool education and school readiness. *American Educational Research Journal*, *41*, 115–157.

Maine Department of Education. (2015). *Pilot training of a kindergarten entry assessment*. Retrieved from https://mainedoenews.net/2015/08/19/pilot-training-of-a-kindergarten-entry-assessment/

Maryland State Board of Education. (2015). *Ready for kindergarten: Maryland's early childhood comprehensive assessment system*. Retrieved from http://marylandpublicschools.org/msde/divisions/child_care/early_learning/docs/KRA2014-15TEchnicalReport.pdf

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*(4), 367–380.

Massachusetts Executive Office of Education. (2016). *Massachusetts Kindergarten Entry Assessment (MKEA)*. Retrieved from http://www.mass.gov/edu/birth-grade-12/early-education-and-care/mkea/

Maxwell, K., Scott-Little, C., Pruette, J., & Taylor, K. (2013). *Kindergarten entry assessment: Smart Start Conference*. Raleigh, NC: North Carolina Department of Public Instruction. Retrieved from http://rtt-elc-k3assessment.ncdpi.wikispaces.net/file/view/KEA%20Smart%20Start%20Presentation%20May%2013%20final.pdf/437991380/KEA%20Smart%20Start%20Presentation%20May%2013%20final.pdf

McAfee, O. D., & Leong, D. J. (2011). *Assessing and guiding young children's development and learning*. Needham Heights, MA: Allyn & Bacon.

McClelland, M. M., Cameron, C. E., Connor, C. M., Farris, C. L., Jewkes, A. M., & Morrison, F. J. (2007). Links between behavioral regulation and preschoolers' literacy, vocabulary, and math skills. *Developmental Psychology*, *43*(4), 947.

McGrew, K. S., & Woodcock, R. W. (2001). *Technical manual: Woodcock-Johnson III*. Rolling Meadows, IL: Riverside.

Meisels, S. J., Jablon, J., Marsden, D. B., Dichtelmiller, M. L., & Dorfman, A. (2001). *The Work Sampling System* (4th ed.). Ann Arbor, MI: Rebus.

Meisels, S. J., Liaw, F. R., 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. http://dx.doi.org/10.1080/1088869100369791

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*(4), 5–8.

Michigan Department of Education. (n.d.). *Kindergarten Entry Assessment (KEA)*. Retrieved from http://www.michigan.gov/mde/0,4615,7-140-22709_65339—,00.html

Minnesota Department of Education. (n.d.). *Minnesota Department of Education announces revised school readiness study: The Kindergarten Entry Profile*. Retrieved from http://education.state.mn.us/mdeprod/idcplg?IdcService=GET_FILE&dDocName=023401&RevisionSelectionMethod=latestReleased&Rendition=primary

Missouri Department of Elementary and Secondary Education. (n.d.). *School Readiness Tool*. Retrieved from https://dese.mo.gov/quality-schools/early-learning/school-readiness-tool

National Center on Quality Teaching and Learning. (2013). *A guide to resources for developing school readiness goals*. Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/tta-system/teaching/docs/state-learning-stds-03-2013.pdf

Nevada Early Childhood Advisory Council. (2016). *Kindergarten Inventory of Development Statewide (KIDS) pilot*. Retrieved from http://nvecac.com/kindergarten-inventory-development-statewide-kids-pilot/

New Jersey Department of Education. (n.d.). *New Jersey Kindergarten Entry Assessment (NJKEA): Information packet 2016-2017*. Retrieved from http://www.state.nj.us/education/ece/rttt/njkea/registration.pdf

New Mexico Public Education Department. (n.d.). *Kindergarten Observational Tool*. Retrieved from http://ped.state.nm.us/ped/LiteracyEarlyChildhoodEd_KOT_index.html

Ohio Department of Education. (2016). *Ohio's Kindergarten Readiness Assessment*. Retrieved from http://education.ohio.gov/Topics/Early-Learning/Kindergarten/Ohios-Kindergarten-Readiness-Assessment

Pavelski-Pyle, R. P. (2002). Best practices in assessing kindergarten readiness. *California School Psychologist*, *7*, 63–73.

Peisner-Feinberg, E. S., Burchinal, M. R., Clifford, R. M., Culkin, M. L., Howes, C., Kagan, S. L., & Yazejian, N. (2001). The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child Development*, *72*, 1534–1553.

Peter, J. P. (1981). Construct validity: A review of basic issues and marketing practices. *Journal of Marketing Research*, *18*(2), 133–145.

Pratt, M. E., McClelland, M. M., Swanson, J., & Lipscomb, S. T. (2016). Family risk profiles and school readiness: A person-centered approach. *Early Childhood Research Quarterly*, *36*, 462–474. doi:10.1016/j.ecresq.2016.01.017

Public Schools of North Carolina. (n.d.). *K-3 formative assessment process*. Retrieved from http://www.dpi.state.nc.us/earlylearning/k3assessment/

Ramey, C. T., & Campbell, F. A. (1991). Poverty, early childhood education, and academic competence: The Abecedarian experiment. In A. Huston (Ed.), *Children in poverty: Child development and public policy* (pp. 190–221). New York, NY: Cambridge University Press.

Ramey, S. L., & Ramey, C. T. (2004). Early learning and school readiness: Can early intervention make a difference? *Merrill-Palmer Quarterly*, *50*(4), 471–491.

Rimm-Kaufman, S. E., Curby, T. W., Grimm, K. J., Nathanson, L., & Brock, L. L. (2009). The contribution of children's self-regulation and classroom quality to children's adaptive behaviors in the kindergarten classroom. *Developmental Psychology*, *45*, 958.

Rimm-Kaufman, S. E., Pianta, R. C., & Cox, M. J. (2000). Teachers' judgments of problems in transition to kindergarten. *Early Childhood Research Quarterly*, *15*, 147–166. doi:10.1016/S0885-2006(00)00049-1

Sadowski, M. (2006). The school readiness gap. *Harvard Education Letter*, *22*(4), 4–7.

Schweinhart, L. J. (1993). *Significant benefits: The High/Scope Perry Preschool study through age 27*. Monographs of the High/Scope Educational Research Foundation, No. 10. Ypsilanti, MI: High/Scope Educational Research Foundation.

Shields, K. A., Cook, K. D., & Greller, S. (2016). *How kindergarten entry assessments are used in public schools and how they correlate with spring assessments*. Retrieved from http://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL_2017182.pdf

Siddens, S., Hubbell, S., & Otto, T. (2013). *Start strong: Ohio's early childhood comprehensive assessment system*. Retrieved from https://ohioedconference.files.wordpress.com/2013/10/kindergartenentryassessment102.pdf

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.

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

Snow, K. L. (2006). Measuring school readiness: Conceptual and practical considerations. *Early Education and Development*, *17*, 7–41.

Snow, K. L. (2011). *Developing kindergarten readiness and other large-scale assessment systems: Necessary considerations in the assessment of young children*. Washington, DC: National Association for the Education of Young Children.

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 State of Washington Office of Superintendent of Public Instruction. Retrieved from http://www.k12.wa.us/WaKIDS/pubdocs/WaKIDS_Report072613.pdf

State of Alabama Department of Education. (2015). Kindergarten entry assessment pilot program. Retrieved from https://www.alsde.edu/sites/memos/Memoranda/FY15-3037.pdf

State of Washington. (2015). *Changes to 2015 WaKIDS whole child assessment*. Retrieved from http://www.k12.wa.us/WaKIDS/Materials/pubdocs/Changesto2015WaKIDSWCA.pdf

Sudkamp, A., Kaiser, J., & Jens, M. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*(3), 743–762.

Teaching Strategies. (2010). *Alignment of the California Desired Results Profile Preschool Instrument with Teaching Strategies GOLD® objectives for development and learning: Birth through kindergarten*. Washington, DC: Author. Retrieved from http://teachingstrategies.com/content/pageDocs/CA-DRDP-GOLD-Alignment-PS-2010.pdf

Teaching Strategies. (2011). *A guide to Teaching Strategies GOLD interrater reliability certification*. Washington, DC: Author.

Teaching Strategies. (2012). *Teaching Strategies GOLD® assessment system: Growth norms technical summary*. Washington, DC: Author.

Teaching Strategies. (2013a). *Teaching Strategies GOLD® assessment system: Concurrent validity*. Washington, DC: Author.

Teaching Strategies. (2013b). *Touring guide*. Washington, DC: Author. Retrieved from http://teachingstrategies.com/content/pageDocs/GOLD-Touring-Guide_5-2013.pdf

U.S. Department of Education. (2013). *U.S. Department of Education awards more than $15.1 million in enhanced assessment grants to develop or improve kindergarten entry assessments*. Washington, DC: U.S. Department of Education Press Office.

Vermont Agency of Education. (2016). *Early childhood education assessment*. Retrieved from http://education.vermont.gov/student-support/early-education/assessment

Walker, H. M., Kavanagh, K., Stiller, B., Golly, A., Severson, H. H., & Feil, E. G. (1998). First step to success an early intervention approach for preventing school antisocial behavior. *Journal of Emotional and Behavioral Disorders*, *6*(2), 66–80.

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*, *2012*, 46º–54.

Wiggins, G. (1990). *The case for authentic assessment*. ERIC Digest.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson Tests of Achievement*. Itasca, IL: Riverside.

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

Wyoming Department of Education. (2016). *WDE kindergarten readiness data collection* (Memorandum No. 2016-029). Retrieved from http://edu.wyoming.gov/downloads/communications/memos/2016/2016-029.pdf

Zill, N., Collins, M., West, J., & Hausken, E. J. (1995). Approaching kindergarten: A look at preschoolers in the United States. *Young Children*, *51*, 35–38.

## Authors

KATHERINE L. MILLER-BAINS is a doctoral student in the Curry School of Education's Research, Statistics, and Evaluation program at the University of Virginia. She is interested in investigating scalable ways to make educational data more useful to practitioners and applying rigorous experimental and quasiexperimental methods to educational program evaluation.

AMANDA M. WILLIFORD is a research associate professor at the Center for Advanced Study of Teaching and Learning within the Curry School of Education at the University of the Virginia. Her research focuses on (a) creating, evaluating, and unpacking early interventions that change children's early education experiences to improve their readiness skills; (b) understanding the proximal processes within the early childhood education context that influence children's development and learning; and (c) applying research to policy and scalable practice.

JACLYN P. RUSSO is a clinical psychology doctoral student within the Clinical and School Psychology Program within the Curry School of Education at the University of Virginia, where she conducts research at the Center for Advanced Study of Teaching and Learning. Her research interests include school readiness, particularly the social and emotional development of young children and how to better help teachers use readiness assessments to guide classroom instruction and interventions.

JAMIE DECOSTER is a senior scientist at the Center for Advanced Study of Teaching and Learning within the Curry School of Education at the University of the Virginia. His research focuses on discovering ways to make the methods practiced by scientists more accurate, flexible, and efficient.

ELIZABETH A. COTTONE is a research scientist at the Center for Advanced Study of Teaching and Learning within the Curry School of Education at the University of the Virginia. Her research interests include investigating pathways from economic disadvantage to poor outcomes for children, understanding attributes of families in poverty through a resilience lens, and contributing to the field's growing recognition of cultural and socioeconomic differences in both early intervention and measurement development with disadvantaged populations.