Authors:
Amanda P. Williford, Jason T. Downer, and Bridget K. Hamre

Acknowledgements:
This study was supported by Elevate Early Childhood through public and private investments. The state appropriated $250,000 from Governor Bob McDonnell’s K-12 Reform Agenda and was matched with $500,000 in private donations from the Batten Educational Achievement Fund of the Hampton Roads Community Foundation. The opinions expressed in this report are those of the authors and do not represent views of the funding agencies. The authors wish to thank the generous school systems, school administrators, teachers, and kindergarten children who participated in this study. In addition we are grateful to Dr. Elizabeth Cottone, Dr. Jamie DeCoster, Elise Rubenstein, Kate Miller-Bains, and Emily Rakosi for their significant contributions to this work.

Correspondence concerning this report should be addressed to Amanda Williford at williford@virginia.edu.
Introduction

Between 33 to 50% of children enter kindergarten behind their peers in at least one early learning domain that is critical for school success.1 Unfortunately, children who enter kindergarten behind their peers rarely catch up; instead, the achievement gap widens over time.2 Children who enter the K-12 school system without being ready are more likely to fall below grade level expectations, to be retained in school, to be placed in special education, to drop out of high school, to not attend college, and to become unemployed.3 These negative outcomes have significant financial costs, such as the costs of special education services and social welfare programs.4

Children from low-income backgrounds enter kindergarten with disproportionally lower school readiness skills (48%) compared to their economically advantaged peers (75%).5 This is known as the school readiness gap.6 In Virginia, 35% of children are from low-income backgrounds and so are at high risk for entering kindergarten behind and never catching up.7

School readiness is the set of skills that children possess when they enter school that prepares them for later school success.8 The key domains of school readiness include language, literacy development, cognition and general knowledge, approaches to learning, physical health (including well-being and motor development), and social and emotional development.9

---

9 See also the Virginia Department of Education's definition of school readiness: http://www.doe.virginia.gov/instruction/early_childhood/school_readiness/va_school_readiness_definition.pdf
These domains are represented in Virginia’s Foundation Blocks for Early Learning which provides a set of comprehensive standards for early learning. School readiness assessments in Virginia, however, have focused only on children’s pre-literacy skills. Through the Early Intervention Reading Initiative (EIRI) enacted in 1997, Virginia schools have resources to assess children’s entry literacy skills, with the vast majority of school districts using the Phonological Awareness Literacy Screening (PALS). Although data from the PALS has provided insight into Virginia kindergartners’ readiness in the literacy domain, we know little about their skills in the other essential school readiness areas.

Despite the importance of understanding the skills children have in these areas when they enter school, accurately assessing young children’s abilities is challenging. Young children’s school readiness skills are often assessed using one or a combination of the following:

1. Checklists/rating scales where teachers rate children on items based upon their knowledge of children during a particular time-frame (e.g. within the last four weeks)
2. Observation/performance-based assessment (also called authentic assessment or work sampling) where teachers observe and document children’s skills during regular classroom activities for a particular set of time (e.g. during the first 6 weeks of school) and then review this documentation to score children on a set of items
3. Direct assessment in which children are administered tasks and/or asked questions by a trained assessor using a standardized protocol (each child is administered the assessment the exact same way) and then scored based upon their response.

Each of these methods has strengths and weaknesses that deserve consideration when choosing how children will be assessed. Although direct assessment is often viewed as the gold standard in measuring children’s school readiness skills, a directly assessed score may not represent how those skills are displayed in the classroom and may not be directly aligned with state standards for student learning. Teacher report measures have been criticized as being potentially biased because they are often strongly related to teacher (rather than child) characteristics. And, children’s demonstration of skills using observation/performance-based assessments depends upon the classroom context (which the teacher creates).

Selecting the best tool also depends upon the purpose and use of the data. For example, a standardized assessment is particularly important if one wants to make comparisons across

11 See PALS site: https://pals.virginia.edu/rd-background.html
children who reside in different classrooms, schools, districts, etc. However, an authentic/performance-based assessment may make sense the most sense if a teacher wants to use the data to individualize her instruction to maximize that child’s learning in the classroom and/or to track children’s growth across the school year.

I. Virginia Kindergarten Readiness Project

The Virginia Kindergarten Readiness Project (VKRP), conducted by Elevate Early Education (E3), in partnership with the University of Virginia’s Center for the Advanced Study of Teaching and Learning (CASTL) and the Virginia Secretary of Education, has the following goals:

• Select an assessment tool that can be used statewide to accurately assess children’s incoming school readiness skills across a range of readiness domains
• Create a snapshot of Virginia’s entering kindergarteners with regard to readiness skills across a comprehensive range of learning domains
• Clearly define school readiness skills gap in Virginia
• Indicate the extent to which estimates of readiness may be different for children from different subgroups (like geographic regions or socio-economic status)
• Inform the implementation of a full-scale statewide comprehensive readiness assessment in the Commonwealth
• Equip E3 with information that can be used to inform public policy and funding decisions in early childhood education

Funding:

This project is funded through public and private investments. The state appropriated $250,000 from Governor Bob McDonnell’s K-12 Reform Agenda and was matched with $500,000 in private donations from the Batten Educational Achievement Fund of the Hampton Roads Community Foundation.

Design:

VKRP is being conducted in two phases:

• **Phase I**: The purpose of this phase of the study is to examine the concurrent validity of Teaching Strategies GOLD when used by teachers to assess children’s school readiness skills in the fall of the kindergarten year. This phase will be completed in Fall 2013.
• **Phase II**: The purpose of this phase is to provide a statewide estimate of entering kindergarten children’s school readiness skills across a range of areas critical for school success. This phase will be completed by the Fall of 2014

*This report focuses on Phase I only.*
Selection of the Kindergarten Readiness Assessment Tool:

Teaching Strategies GOLD (TS GOLD)\(^\text{16}\) is an observation-based (sometimes called performance-based), teacher-administered readiness assessment that is used during regular classroom activities. Teachers observe children demonstrating skills during typical instruction and document evidence of their progress in multiple domains of school readiness skills. Covered skill areas include mathematics, literacy, cognitive, language, physical, and social-emotional development, and teachers can track children’s growth in these readiness domains over time.

TS GOLD was chosen as the initial assessment tool for several reasons.

1. It is a teacher-administered tool based upon authentic assessment. Teachers observe and collect data on children as they engage in during regular classroom day. These procedures are purported to allow teachers to accurately assess children’s skills.

2. It assesses a broad range of school readiness skills.

3. It can be customized to align with state-specific learning standards.

4. It was ready for statewide scalability at the time the VKRP was initiated. More than 20 state departments of education for preschool, prekindergarten, and/or kindergarten entry assessment currently use TS GOLD.

5. The Virginia Early Childhood Association (VECF) had previously chosen TS GOLD for a feasibility study to answer questions about the extent to which GOLD is practical for teachers to administer and use as a tool to inform practice, differentiate instruction, and assess children’s skill growth over time (questions this study was not designed to answer).

Measures are only useful to the extent that they consistently and accurately assess the constructs they are meant to represent. The publishers have examined the measurement properties of TS GOLD and report the measure assesses children consistently\(^\text{17}\), and they recently reported evidence of concurrent validity in a sample of preschool children.\(^\text{18}\) However, there are no independent studies that examine the extent to which children’s estimates of kindergarten readiness using TS GOLD are related to direct measurements of children’s kindergarten readiness. Evidence of such validity is critical for having confidence to say whether TS GOLD’s reports of kindergarten readiness are accurate representations of children’s actual skills.

Thus, in Phase 1 of the VKRP we concurrently examined the validity of TS GOLD in a sample of Virginia kindergarten children by comparing children’s school readiness skills reported by TS GOLD to reports provided by well-established direct assessments of children’s school readiness skills collected at the same time.

---


II. Methodology

Sample:
The sample included 1086 kindergarten students enrolled within 58 kindergarten classrooms and 16 elementary schools across four diverse districts in the state. The sample was not chosen to directly represent the state, but instead to contain teachers and children with a diversity of demographic characteristics (see Appendix, Tables 1 and 2). Of the 1086 students in the sample, 18% of students were non-white and 8% were identified as having special education needs. We were unable to obtain free- and reduced-lunch information at the child level for all students. However, school-level information obtained from VDOE from 2012-13 indicated that children came from diverse SES backgrounds (with an average of 44% of students being eligible for free- or reduced lunch). Fifty-two of the 58 classrooms were randomly selected to participate in the validity assessments. Within each classroom, 10 students were randomly selected from the classroom roster to be directly assessed.

Participating teachers had a wide range of experience. On average, teachers had been teaching for 14 years. One quarter of teachers had been in the classroom for four years or less and around 30% had been teaching for 20 years or more. All teachers had earned a Bachelor’s degree or higher, with more than half of teachers (56.9%) possessing a Master’s degree. The vast majority of sampled teachers were white (91.4%) and, much like the nation as a whole, almost all were female (98.3%).

Procedures:

Teacher Assessments. Teachers observed and rated their students’ readiness skills using TS GOLD. Prior to assessing their classrooms, all teachers participated in an onsite two-day training provided by TS GOLD during which they were instructed on how to document and rate children’s readiness skills using the TS GOLD criteria. Per recommendations from the developers of TS GOLD, teachers were encouraged but not required to complete an interrater reliability task for all six developmental domains. This process involves teachers reviewing sample portfolios/documentation and then rating children’s skills based upon review of the material. If 80% of a teacher’s item ratings in a developmental area fell within one point of the master score, then the teacher was certified as reliable for that domain. Almost all teachers started the reliability certification process, 50% completed the tasks, and 28% of teachers achieved reliability certification for all domains.

Following training, teachers were then asked to observe and document children’s readiness skills for four weeks and to subsequently assess children’s skills during a two-week fall assessment window. The TS fall assessment window occurred 7 – 11 weeks from the start of student instruction. For all districts, the GOLD fall assessment window coincided with PALS fall administration window. In a post-assessment survey, teachers reported a wide range of time taken

---

19 Power analyses indicated a sample of 50 classrooms and 500 children was sufficient to detect modest correlations between TS GOLD and the direct assessments. However, districts were given the choice of allowing all their teachers to use TS GOLD regardless of whether their classroom was randomly chosen for the direct assessments. Timing and funding constraints did not allow for us to conduct direct assessments within all 58 classrooms.


22 78% of participating teachers (45 out of 58) completed an online satisfaction survey within 1-2 weeks of finalizing the TS GOLD checkpoint. Teachers responded to the questions: “Overall, how much time did you spend per child preparing TS GOLD
to complete the GOLD assessment in their classroom. For the majority of teachers, the TS GOLD documentation took an average of 1 – 2 hours per child (20 to 40 hours per classroom), while Fall Checkpoint finalization took an additional 30 minutes to 1 hour per child (10 to 20 hours per classroom). Thus, the total time per child was 1.5 to 3 hours or 30 to 60 hours per classroom.

**Validity Assessments.** Eleven independent data collectors were hired and completed a two-day training at University of Virginia. Following the training, data collectors were required to practice the assessment administration, and all data collectors were supervised in their initial administration in order to ensure fidelity to the direct assessment procedures. Direct assessments were conducted within 6 – 10 weeks from the start of student instruction and overlapped with the TS GOLD fall assessment window. The validity assessments took between 20 and 45 minutes per child to complete (7 to 15 hours to assess all children in each classroom).

**Measures:**

**Teaching Strategies Gold.** As described above, TS GOLD was selected as the teacher-administered readiness assessment measure. Teachers rated children on all items which included the following areas: literacy (12 items), mathematics (7 items), language, 8 (items), social-emotional (9 items), cognitive (10 items), science and technology (5 items), social studies (4 items), arts (4 items) and English language acquisition (2 items). For this study, we used the five areas (literacy through cognitive) where items are rated on a 9-point color banded scale as these areas have raw and scaled scores and cut-offs for the Widely Held Expectations (described below).

**Direct Assessments.** The Woodcock-Johnson-III Psychoeducational Battery \(^{23}\) (WJ-III) is a widely used, individually administered assessment battery that measures general cognitive abilities and achievement in individuals from age two through adulthood. Twelve achievement subtests can be used with preschoolers, all demonstrating good internal reliability and validity. We used one subscale to measure children's language skills: Picture Vocabulary which measures expressive vocabulary. We used two subscales to measure children's literacy skills: Letter Word Identification which measures children's letter and word identification and Word Attack which measures phonic and structural analysis of words. These subscales combine to form a Broad Reading Composite score which was used in the analyses. We used two subscales to measure children's mathematics skills: Applied Problems which measures analytical and problem solving skills and Quantitative Concepts which measures math knowledge of concepts, symbols and vocabulary.

Children's Self-Regulation skills were assessed using two measures. Children's inhibitory control was assessed using the Pencil Tap Test.\(^{24}\) Percent of correct responses is computed as the child's score. Children's behavioral self-regulation, including inhibitory control, working memory and attention focusing was assessed using the Head Toes Knees Shoulders Task.\(^{25}\) Both tests have been widely used in recent developmental and early education research and show good reliability and concurrent and predictive validity.\(^{26}\)

---

III. Results

Table 3 in the Appendix provides descriptive statistics for the variables used in this study. Below we describe the main results of the study.

**Question 1: Do TS GOLD and the direct assessments provide a similar picture regarding individual children’s school readiness skills?**

Answer: Children's literacy and math scores as assessed by TS GOLD were similar to their scores in direct assessments in those domains. Children's language and cognition skills as assessed by TS GOLD were not as similar to their direct assessments of language and self-regulation skills respectively. Figure 1 illustrates the alignment between the TS GOLD subscales and the corresponding direct assessment.

Implication: These results provide evidence that TS GOLD accurately measures (i.e., concurrent, construct validity) children's literacy and math skills.

![Figure 1. Alignment between TS GOLD and direct assessments](image)

Note. The vertical axis represents the bivariate correlations between TS GOLD and the direct assessments (degree of alignment). You want the diamonds that are the same color (the corresponding direct assessment) as the TS GOLD domain to be above .50.

org/10.1016/j.ecresq.2007.01.002

Explanation and Evidence:

Pearson correlation coefficients were used to estimate the associations of TS GOLD with direct assessments (See Appendix Table 4). The strength of these associations varied based on readiness domains. In other samples the associations between teacher report of children's skills (such as TS GOLD) and direct assessments of similar constructs typically range from .40 to .60.27 The strength of the relations is comparable to prior research relating other teacher report with direct assessments of similar constructs for math (.64) and literacy (.53) but were lower for language (.36) and cognition (with self-regulation; .27 and .31).

Question 2: How well does TS GOLD capture differences in readiness skills across developmental domains for individual children (e.g., domain specificity)?

Answer: We found that teachers scored children's skills much more similarly across school readiness domains using TS GOLD than was evident when children were measured using direct assessments. For example, when teachers rated children 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,28 we would expect less consistency in children's skills across readiness domains than was demonstrated when children were assessed by teachers using TS GOLD. Figure 2 illustrates this result by providing an example child profile. In addition, we found the within-domain associations between TS GOLD and the direct assessments (e.g., TS GOLD literacy with the direct assessment of literacy) were not stronger than the cross-domain associations (e.g., TS GOLD literacy with the direct assessment of math).

Implication: TS GOLD provides data on children across a comprehensive range of readiness domains. The value of having a comprehensive assessment is that it will show variability in children's skill levels across different readiness domains when they exist. However, children tend to look the same across readiness domains when assessed using TS GOLD. Thus, TS GOLD does not appear to be sensitive enough to determine which children might be performing differently across readiness domains.


Explanation and Evidence:

The within assessment associations (using Pearson correlation coefficients) are presented in the Appendix Tables 4 and 5. Consistent with prior research\footnote{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. \textit{Language, Speech, and Hearing Services in Schools}, 40, 161-173. \url{http://dx.doi.org/10.1044/0161-1461(2009/07-0099)}}\textsuperscript{29}, the associations among the school readiness domains as measured through direct assessment were positive and moderate in strength with a median correlation of .40 and ranged from low (.19 for literacy with self-regulation) to moderate (.61 for literacy with math). Thus, a child’s score tended to be variable across readiness domains when measured via direct assessment. In contrast, the associations among the school readiness domains as measured by TS GOLD were positive and high in strength with a median correlation of .76 and ranged from moderate (.59 for literacy with physical) to high (.85 for language with cognitive).

Another important aspect of specificity is that scores measuring the same or similar domains are associated more highly with each other than scores that measure different domains (i.e., discriminant validity). We compared the associations of TS GOLD scores with direct assessments measuring the same readiness domain (e.g., the scores that both measure math) with the associations of TS GOLD with direct assessments measuring different readiness domains (e.g., a score of math and a score of language).

In most cases, the within-domain associations were not stronger than the cross-domain associations.
associations. Appendix Table 6 provides the full correlation matrix, which is illustrated in Figure 3. This guide can be used to interpret Figure 3: the diamond that is the same color (direct assessment measuring the same readiness domain) as the TS GOLD domain should be above .50 AND be higher when compared to the other colored diamonds (direct assessments measuring other readiness domains).

**Figure 3. Domain specificity between TS GOLD and direct assessments**

TS GOLD Math was the only domain that demonstrated specificity with regard to its associations with the direct assessments. The results for each TS GOLD domain are described below:

- **TS GOLD Literacy**: The correlation of TS GOLD literacy (first column) with the direct assessment of literacy (blue diamond, .64) was the same magnitude as the correlation of TS GOLD literacy with the direct assessment in math (red diamond, .63).

- **TS GOLD Language**: The correlation of TS GOLD language (second column) with the direct assessment of language (purple diamond) was lower than the correlations of TS GOLD language with direct assessments in literacy (blue diamond) and math (red diamond).

- **TS GOLD Math**: The correlation of TS GOLD math (third column) with the direct assessment of math (red column) was higher than the correlations of TS GOLD math with direct assessments in the other areas. Thus, TS Math shows domain specificity with regard to its associations with the direct assessments.

- **TS GOLD Cognitive**: The correlation of TS GOLD cognitive (fourth column) with the direct assessments of self-regulation (orange diamonds) was lower than the correlations of TS GOLD cognitive with direct assessments in literacy (blue diamond) and math (red diamond).
**Question 3: Does TS GOLD measure children from different backgrounds equitably?**

**Answer:** Yes, TS GOLD measures children from different backgrounds equitably.

**Implication:** There is no evidence to indicate that using TS GOLD would be more appropriate for some children but not others.

**Explanation and Evidence:**

If TS GOLD measures children from different backgrounds equitably, we would expect the associations of TS GOLD with the direct assessments to be the same for groups of children that have different demographic characteristics. Thus, we examined whether the associations of TS GOLD with direct assessments of similar constructs varied by gender, child race, geographic region, or child socio-demographic data. The results of these analyses suggest that the associations of TS GOLD with the corresponding direct assessments were consistent across child gender, race, SES, and geographic location.

**Question 4: Does TS GOLD differentiate children’s readiness skills within a classroom?**

**Answer:** Children within a classroom look much more similar with regard to their level of readiness skills when measured via TS GOLD as compared to direct assessments. This result is illustrated in Figure 4 through an example child profile.

**Implication:** These results suggest that TS GOLD is more influenced by teacher characteristics than direct assessments. This is problematic, given that these measures are intended to assess skill differences between children.

**Explanation and Evidence:**

Because there are natural similarities among children going to the same school (e.g., they may come from similar neighborhoods, they share the same classroom context), we do expect to see somewhat similar ratings for children in the same classroom. However, if children’s readiness scores within a classroom are much more similar when measured via TS GOLD than their scores using direct assessments (which are independent of teacher), it would suggest that there may be some teacher bias in the TS GOLD ratings. The similarity of children’s scores within classrooms was estimated using an intra-class correlation coefficient (ICC). ICCs range from 0 to 1, with higher scores indicating greater within-classroom similarity. There was much higher similarity within classrooms based on TS GOLD (ICCs ranged from .28 [literacy] to .62 [physical] with a median ICC of .45) than for direct assessments (ICC’s ranged from .01 [self-regulation] to .04 [math] with a median ICC of .03).

Publishers of TS GOLD claim that teachers can use their tool to objectively measure children’s skills. The objectivity in teachers’ scores is supposed to be the result of comprehensive training on how to conduct and document ongoing observations of each child’s demonstrated skills across the full range of readiness domains during regular classroom activities. We examined the ICCs for TS GOLD and the direct assessments and did not find strong evidence for this assertion. Instead, our results suggest that TS GOLD is likely subject to teacher bias and that much of the variability in scores may reflect teacher perceptions rather than children's skill levels.
Question 5: Do TS Gold and the direct assessments provide a similar picture in terms of the proportion of children who are “school ready”? 

Answer: TS GOLD identified a large percentage of children as not meeting expectations for kindergarten. However, many children who were identified by the TS GOLD Widely Held Expectation cut-offs as “not meeting expectations” fell at or near the national average for their age when measured via direct assessment.

Implication:
One of the major purposes in implementing a kindergarten readiness assessment is to identify children who do not have the skills needed to be successful in kindergarten. Using TS GOLD will likely identify many children at or near the national average with regard to skill level as “not ready”. This suggests that TS GOLD is setting a higher benchmark for readiness than the standard typically used by schools.

Explanation and Evidence:
TS GOLD categorizes children within each readiness domain as “not meeting”, “meeting”, or “exceeding expectations”. This categorization is called the Widely Held Expectations (WHE). The cut-off points were determined by TS Gold developers based upon developmental and early education research literature. We wanted to know where children who were categorized as “not meeting expectations” scored on the direct assessments that have nationally
normed standard scores. For example, the WJ-III is a nationally normed achievement test that has an average standard score of 100 and a standard deviation of 15. Thus, a score of 100 is at the 50th percentile and is the average or typical score for a child of that age. We related the TS GOLD cut-off for not meeting expectations to the standard score on the WJ for each readiness domain.

TS GOLD WHE indicated that a large percentage of children were not meeting expectations for kindergarten (68% for math, 65% for language, and 30% for literacy). The cut-offs for not meeting expectations corresponded to a standard score of 100 for literacy (50th percentile), 103 (58th percentile) for language, and 104 (61st percentile) for math. Specifically, children must reach at least the 50th percentile or higher in terms of their skill level in these readiness domains to be identified as ready by TS GOLD.

IV. Limitations

It is important to acknowledge the limitations of this study. The sample for this study is relatively small which limits our confidence in broadly generalizing these results. Children came from diverse backgrounds. However, the teachers were mostly White with varied levels of experience. Additionally, this was the teachers' first experience using TS GOLD. It is possible that the accuracy of teachers’ ratings of students using TS GOLD would change after they had used the measures for an extended period of time. Finally, although all teachers completed the required, in-person, 2-day training before beginning to use TS GOLD, we did not require them to complete TS GOLD’s formal certification process as this is not required by the publishers for teachers to use the measure. Completing this process may have improved teachers' accuracy in assessing children's readiness skills.

V. Summary and Recommendations

Teaching Strategies Gold is an observation-based, authentic assessment tool that measures children's school readiness skills across a comprehensive range of readiness domains. The information obtained about children's school readiness skills as measured by teachers using TS GOLD was compared to information obtained about children's school readiness skills measured by independent data collectors using well-established direct assessments in order to examine the accuracy of TS GOLD. We found that:

- TS GOLD assessments of literacy and math skills were well aligned with the direct assessment of those skills, which shows some evidence of concurrent, construct validity.

- TS GOLD appears to measure children from different backgrounds equitably.

- TS GOLD did not show good specificity in children's scores across readiness domains. Children looked much more similar when assessed by TS GOLD compared to the direct assessments suggesting that TS GOLD is not sensitive enough to determine how children are performing differentially across readiness domains.

- Children within a classroom looked much more similar with regard to their level of readiness skills when measured via TS GOLD as compared to the direct assessments suggesting that TS GOLD is subject to teacher bias.
• TS GOLD sets a much higher benchmark for determining readiness than the standard typically used by schools and early education research. Using the TS GOLD Widely Held Expectations will likely identify many children at or near the national average with regard to skill level as “not ready”.

Recommendations
Measuring the readiness competencies of incoming kindergarten children is a challenging task and no assessment will measure children’s skills without some error. For example, when used in high stakes situations (such as demonstrating that children learned during the year) teacher administered direct assessments have been found to show just as much teacher associated error as using work sampling systems or teacher report. ³⁰ Common options for measuring children’s school readiness skills include checklists/rating scale, observation/performance-based assessment (also called authentic assessment or work sampling), and direct assessment (conducted either by the child’s teacher or an independent assessor). Each of these assessment methods has strengths and weaknesses and making the most appropriate choice for the type of assessment as well as the method of administration should be based upon the purpose of the data—how the data are intended to be used.

This is not a simple decision and there is not a single right answer. This is evident in the fact that states have made various decisions about how to move forward with kindergarten readiness assessments. We reviewed what states are currently implementing and developing in terms of kindergarten readiness assessments, and a number of states are using TS GOLD (many are in the pilot phase). It is also common for states to have developed their own measures. For the most part, these state-originated assessments use the same methodology as TS GOLD (teachers make repeated observations of children over time and collect documentation of children’s performance across readiness domains and subsequently rate children during a particular time-frame). We expect that these assessments will perform similarly to TS GOLD if analyzed as we have done in this report. There are also a number of states that have recently received or have requested state and federal (including Race to the Top funding) and are testing out innovative methods for assessing children’s kindergarten readiness skills. Below are two examples:

• Oregon – Through a combined effort of the Early Learning Council and the state DOE, Oregon has developed and piloted their own statewide kindergarten assessment which uses a composite assessment including teacher direct assessments of literacy and math using the Easy CBM literacy and math assessments (a direct assessment available through Riverside) and teachers’ ratings of children’s social skills and self-regulation using the Child Behavior Rating Scale.³¹,³² It should be noted that they considered using a modified version of TS GOLD and chose the above method as it was better aligned with current assessment practices and also involved lower financial cost and less training and admin-

Ohio & Maryland – These states have combined efforts to develop and field-test a new comprehensive readiness assessment. Their kindergarten entry assessment consists of observation rubrics, performance tasks involving a scripted scenario, and selected response with manipulatives. They are concurrently developing technological supports for teachers to assist with data collection/organization and provide professional development opportunities.

We recommend that Virginia seek out details from such states to understand exactly what these states have developed or adopted, why they have made these choices, and consider whether portions of some state’s kindergarten assessments might be suitable to combine with the kindergarten fall PALS data in order to obtain a more comprehensive measure of children’s readiness skills in Virginia.

Based upon E3’s priority for a kindergarten readiness assessment that measures children’s skills at the beginning of kindergarten across multiple domains of learning and to be able to compare the level of children’s readiness skills across child characteristics such as age, gender, race and socio-economic backgrounds as well as across classrooms, districts, and regions across the state, it is recommended that E3 consider piloting additional options beyond TS GOLD for the purposes of providing a snapshot of school readiness in Virginia. This recommendation is based upon the results of this study that indicate TS GOLD does not seem to be differentiating across readiness domains within individual children, appears subject to teacher bias, and appears to over-identify children as being “not ready”.

Below we provide several examples of how to move ahead with a more comprehensive kindergarten readiness assessment in Virginia based upon two different data usage scenarios.

1. Data will be used solely to provide a snapshot of children’s school readiness skills as they enter kindergarten in Virginia.
   A reasonable approach would be to use a fairly comprehensive assessment protocol (direct and teacher-report) administered to a random sample (stratified by region) to only a portion of children that would be representative of children in Virginia as a whole. For example, since 2002 the Minnesota DOE assesses children’s kindergarten readiness skills based on a 10% random sample of schools stratified by region. This approach would require a relatively small number of data collectors to administer the assessment to a relatively small number of children. An example of the assessment protocol would be: WJ-III Achievement to measure literacy, language, math, and knowledge, Head Toes Knees Shoulder to assess Self-Regulation and a teacher report such as the Social Skills Rating System to assess children’s social-emotional skills. However, there are other direct assessments and rating scales that could also be used. In this option, we envision that PALS administration would continue as currently implemented. However, trained data collectors would assess children’s literacy so that results would be comparable to the other academic domains.

34 http://ohioedconference.files.wordpress.com/2013/10/kindergartenentryassessment1028.pdf
2. Kindergarten readiness data collected at the beginning of the year will be used by kindergarten teachers to inform classroom instruction.

In this approach, it makes sense to assess all children. And, one would want the scores to be easily accessible and clearly understood by teachers, as well as readily linked to strategies the teacher could use to change instruction to maximize learning. There are multiple possibilities for accomplishing these goals, two of which are noted below:

a) Kindergarten teachers could administer a comprehensive assessment protocol similar to what is being done in Oregon. We would recommend complementing PALS, which is already in place, with a direct assessment of math (e.g., the Easy CBM for math or Early Math Diagnostic Assessment\(^{36}\)) and a teacher rating of children’s social-emotional skills with national norms.

b) Training protocols and reliability assessments for TS GOLD could be further developed and piloted to discern whether it is possible to increase teachers’ capacity to differentiate among children and between readiness skills within children when using this authentic assessment tool.

Once an option for moving forward is chosen, the CASTL research team will consult with colleagues who can provide detailed information and further recommendations about measure selection and going to scale statewide. These individuals could include but not be limited to: Thomas Schultz – a nationally recognized expert in early childhood assessment who consults with several other states; Gail Joseph who led kindergarten readiness assessment efforts in Washington State; and/or Kathy Thornburg who led readiness assessment efforts in Missouri. We anticipate being able to recommend alternative measures by mid-February and could do some field testing in the early spring to obtain estimates on time and training.

Next fall, we propose to pilot the chosen assessment battery in order to obtain a statewide estimate of entering kindergarten children's school readiness skills across a range of areas critical for school success. The exact method of recruitment will depend upon how one envisions assessing children in Virginia going forward (i.e., through a random sample or assessing all children). However, in either case, we will sample in a way that will be representative of the state’s population. In order to guarantee geographic representation, we will block by the eight regions in Virginia and select the sample from each region in proportion to the total state enrollment in that region. Recruitment will need to begin by the end of January in order to have sufficient time to partner with selected schools.

Finally, it is critically important for all relevant stakeholders (e.g., E3, VECF, UVA, VDOE) to work together to ensure that the data and findings from multiple pilots using different assessment measures will be used to inform the development of a kindergarten readiness assessment approach that can be recommended to the legislature in January of 2015.

--

### Table 1. Student Characteristics

<table>
<thead>
<tr>
<th></th>
<th><strong>Total Sample</strong></th>
<th></th>
<th><strong>Validity Subsample</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>n</em></td>
<td>%</td>
<td><em>n</em></td>
<td>%</td>
</tr>
<tr>
<td>Number of Students</td>
<td>1086</td>
<td>--</td>
<td>522</td>
<td>--</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>578</td>
<td>53.2%</td>
<td>262</td>
<td>50.2%</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>790</td>
<td>72.7%</td>
<td>383</td>
<td>73.4%</td>
</tr>
<tr>
<td>Black</td>
<td>178</td>
<td>16.4%</td>
<td>83</td>
<td>15.9%</td>
</tr>
<tr>
<td>Asian</td>
<td>20</td>
<td>1.8%</td>
<td>11</td>
<td>2.1%</td>
</tr>
<tr>
<td>Multi-Racial</td>
<td>48</td>
<td>4.4%</td>
<td>22</td>
<td>4.2%</td>
</tr>
<tr>
<td>Other/Unknown Race</td>
<td>50</td>
<td>4.6%</td>
<td>23</td>
<td>4.4%</td>
</tr>
<tr>
<td>Free Reduced Lunch Eligibility</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>227</td>
<td>20.9%</td>
<td>95</td>
<td>18.2%</td>
</tr>
<tr>
<td>No</td>
<td>208</td>
<td>19.2%</td>
<td>85</td>
<td>16.3%</td>
</tr>
<tr>
<td>Not Reported</td>
<td>651</td>
<td>59.9%</td>
<td>180</td>
<td>65.5%</td>
</tr>
<tr>
<td>Special Education Status</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has IEP</td>
<td>89</td>
<td>8.2%</td>
<td>35</td>
<td>6.7%</td>
</tr>
<tr>
<td>Does not have IEP</td>
<td>997</td>
<td>91.8%</td>
<td>487</td>
<td>93.3%</td>
</tr>
</tbody>
</table>

No categories significant at p<.05
Table 2. Teacher Characteristics

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>58</td>
<td>--</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>57</td>
<td>98.3%</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>53</td>
<td>91.4%</td>
</tr>
<tr>
<td>Black</td>
<td>3</td>
<td>5.2%</td>
</tr>
<tr>
<td>Multi-Racial</td>
<td>2</td>
<td>3.4%</td>
</tr>
<tr>
<td>Highest Degree Achieved</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bachelors</td>
<td>25</td>
<td>43.1%</td>
</tr>
<tr>
<td>Masters</td>
<td>33</td>
<td>56.9%</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age, Years</td>
<td>41</td>
<td>11</td>
</tr>
<tr>
<td>Years of Experience</td>
<td>14</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 3. Descriptive Statistics

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOLD Literacy</td>
<td>1064</td>
<td>59.28</td>
<td>15.72</td>
<td>3.00 – 97.00</td>
</tr>
<tr>
<td>GOLD Language</td>
<td>1067</td>
<td>48.85</td>
<td>8.75</td>
<td>0.00 – 70.00</td>
</tr>
<tr>
<td>GOLD Math</td>
<td>1064</td>
<td>36.64</td>
<td>8.24</td>
<td>0.00 – 57.00</td>
</tr>
<tr>
<td>GOLD Cognitive</td>
<td>1063</td>
<td>53.98</td>
<td>10.82</td>
<td>7.00 – 86.00</td>
</tr>
<tr>
<td>GOLD Social/Emotional</td>
<td>1067</td>
<td>52.88</td>
<td>10.50</td>
<td>5.00 – 79.00</td>
</tr>
<tr>
<td>GOLD Physical</td>
<td>1064</td>
<td>33.13</td>
<td>4.95</td>
<td>8.00 – 43.00</td>
</tr>
<tr>
<td>WJᵃ: Basic Reading</td>
<td>521</td>
<td>392.13</td>
<td>26.52</td>
<td>328.00 – 507.00</td>
</tr>
<tr>
<td>WJ: Picture Vocabulary</td>
<td>520</td>
<td>473.66</td>
<td>9.35</td>
<td>440.00 – 505.00</td>
</tr>
<tr>
<td>WJ: Math Reasoning</td>
<td>516</td>
<td>432.80</td>
<td>14.58</td>
<td>381.00 – 473.00</td>
</tr>
<tr>
<td>Pencil Tapᵇ</td>
<td>522</td>
<td>13.92</td>
<td>3.41</td>
<td>0.00 – 16.00</td>
</tr>
<tr>
<td>Head-Toes-Knees-Shouldersᶜ</td>
<td>522</td>
<td>26.08</td>
<td>12.54</td>
<td>0.00 – 40.00</td>
</tr>
</tbody>
</table>

a. WJ=Woodcock Johnson III W Score
b. Raw total score
c. Raw combined score for Parts 1 & 2
### Table 4. Correlations between GOLD Dimensions and Direct Assessments

<table>
<thead>
<tr>
<th></th>
<th>GOLD</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Literacy</td>
<td>Language</td>
<td>Math</td>
<td>Cognitive</td>
</tr>
<tr>
<td>WJ: Basic Reading</td>
<td><strong>.640</strong></td>
<td>.403**</td>
<td>.416**</td>
<td>.364**</td>
</tr>
<tr>
<td>WJ: Picture Vocabulary</td>
<td><strong>.439</strong></td>
<td><strong>.364</strong></td>
<td>.320**</td>
<td>.317**</td>
</tr>
<tr>
<td>WJ: Math Reasoning</td>
<td><strong>.631</strong></td>
<td>.472**</td>
<td><strong>.532</strong></td>
<td>.464**</td>
</tr>
<tr>
<td>Pencil Tap</td>
<td><strong>.259</strong></td>
<td>.265**</td>
<td>.223**</td>
<td><strong>.265</strong></td>
</tr>
<tr>
<td>Head-Toes-Knees-Shoulders</td>
<td><strong>.369</strong></td>
<td>.305**</td>
<td>.297**</td>
<td><strong>.305</strong></td>
</tr>
</tbody>
</table>

*p <.005, **p <.001
Correlations between conceptually related GOLD Dimensions and Direct Assessments are in bold

### Table 5. Within TS GOLD Dimensions Correlations

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Literacy</td>
<td>--</td>
<td>.760**</td>
<td>.761**</td>
<td>.790**</td>
<td>.617**</td>
<td>.586**</td>
</tr>
<tr>
<td>2. Language</td>
<td></td>
<td>--</td>
<td>.795**</td>
<td>.854**</td>
<td>.799**</td>
<td>.720**</td>
</tr>
<tr>
<td>3. Math</td>
<td></td>
<td></td>
<td>--</td>
<td>.792**</td>
<td>.633**</td>
<td>.598**</td>
</tr>
<tr>
<td>4. Cognitive</td>
<td></td>
<td></td>
<td></td>
<td>--</td>
<td>.806**</td>
<td>.680**</td>
</tr>
<tr>
<td>5. Social/Emotional</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>--</td>
<td>.708**</td>
</tr>
<tr>
<td>6. Physical</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>--</td>
</tr>
</tbody>
</table>

**p <.001

### Table 6. Within Direct Assessments Correlations

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. WJᵃ: Basic Reading</td>
<td>--</td>
<td>.475**</td>
<td>.610**</td>
<td>.173**</td>
<td>.300**</td>
</tr>
<tr>
<td>2. WJ: Picture Vocabulary</td>
<td></td>
<td>--</td>
<td>.510**</td>
<td>.230**</td>
<td>.405**</td>
</tr>
<tr>
<td>3. WJ: Math Reasoning</td>
<td></td>
<td></td>
<td>--</td>
<td>.534**</td>
<td>.367**</td>
</tr>
<tr>
<td>4. Pencil Tapᵇ</td>
<td></td>
<td></td>
<td></td>
<td>--</td>
<td>.404**</td>
</tr>
<tr>
<td>5. Head-Toes-Knees-Shouldersᶜ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>--</td>
</tr>
</tbody>
</table>

**p <.001

ᵃ. WJ = Woodcock Johnson III W Score  
b. Raw total score  
c. Raw combined score for Parts 1 & 2