NIEER
NATIONAL INSTITUTE FOR
EARLY EDUCATION RESEARCH

TECHNICAL REPORT: KINDERGARTEN EARLY LEARNING SCALE

September 2014

*Shannon Riley-Ayers,
Kwanghee Jung, &
Jorie Quinn,
The National
Institute for Early
Education Research*

Table of Contents

About the Authors

Shannon Riley-Ayers, Ph.D. Dr. Riley-Ayers is an Assistant Research Professor at The National Institute for Early Education Research (NIEER) at Rutgers University. Dr. Riley-Ayers conducts research at NIEER on issues related to literacy, performance-based assessment, and professional development–often working with teachers and early childhood leaders. She is also on staff with the Center for Enhancing Early Learning Outcomes (CEELO), a federally funded comprehensive center that provides technical assistance to state agencies around early childhood issues. Look for her recent CEELO policy report entitled Formative Assessment Guidance for Early Childhood Policymakers. She is co-author with Dorothy Strickland of the policy brief Early Literacy: Policy and Practice in the Early Years (NIEER) and the book Literacy Leadership in Early Childhood: The Essential Guide (Teachers College Press) and has several other publications on literacy in early childhood. She is first author of the Early Learning Scale Preschool (Lakeshore Learning Materials), a comprehensive performance-based assessment system for preschool children. She also led the validation study of this instrument and continues to evaluate its implementation and use in the field. Dr. Riley-Ayers co-leads several additional research projects at NIEER, including the development and validation of an early childhood quality teacher survey and an alignment study of kindergarten entry assessments in San Antonio. Before joining NIEER, Dr. Riley-Ayers was co-director of the Office of Early Literacy at the New Jersey Department of Education and was instrumental in developing and implementing the New Jersey Early Literacy Initiative. She is a certified teacher and reading specialist, with several years of experience in public school classrooms. She holds a M.Ed. in language and literacy and a Ph.D. in educational psychology from The Pennsylvania State University.

Kwanghee Jung, Ph.D. Dr. Jung has been Principal Investigator on evaluation studies of state funded preschool programs, including the Arkansas Better Chance Pre-K, multi-state evaluation of Acelero Head Start Program, and New Jersey Abbott Preschool Programs. Using advanced statistical models for quasi-experimental designs, such as regression-discontinuity design (RDD) and other matched-group designs, she examined the effects of the programs on children's learning and school-readiness in eight states (AR, CA, MI, NM, NJ, OK, SC, and WV). She has also been involved in several randomized trial studies that examine the effects of various pre-k program features, such as reduced class size, extended day length, dual language environment, and use of the Tools of the Mind curriculum. Dr. Jung was the lead analyst for validation studies of the CASEBA, PRISM, and ELS. She also provided psychometric analysis on a study which looked at the reliability and validity of children's performance task measures in evaluating preschool and kindergarten students' literacy and math skills.

Jorie Quinn, Ed.D. Dr. Quinn is a research coordinator at the National Institute for Early Education Research (NIEER) and has overseen information and data collection on various studies. She has provided coaching and professional development workshops to early childhood teachers, and has worked in Early Childhood Research at NIEER for almost two years. Before joining NIEER she was the Associate Director for Early Childhood Education at Liberty Science Center (LSC) for almost five years. While at LSC, Dr. Quinn collaborated with the New Jersey

Department of Education to deliver STEAM-based professional development workshops. There, she designed and developed their early childhood program for children, parents and caregivers, and teachers. She was recognized for her work in early childhood when Liberty Science Center was voted one of the Ten Best Science Centers in the country by Parent's Magazine in 2008. Before joining LSC, Dr. Quinn was a preschool teacher for two years, adjunct professor at Fairleigh Dickinson University for eight years, and Technology Director for a K-6 school district for two years. Additionally, she provided coaching, workgroups, and workshops for preschool and kindergarten teachers in four schools in northern New Jersey for three years. Dr. Quinn is a certified teacher for K-8 and holds a supervisor certificate. She earned a M.Ed. from Fairleigh Dickinson University in Education and an Ed.D. in Educational Leadership with a specialization in Curriculum and Development from the University of Phoenix.

Grateful acknowledgement is made to Monica DellaMea and Clayton Burch of West Virginia Department of Education for their support of this project.

Correspondence regarding this report should be addressed to Shannon Riley-Ayers at the National Institute for Early Education Research. Email: sayers@nieer.org.

Introduction

In the age of accountability, data collection seems to be in vogue. Data are now routinely collected nationwide on children, classrooms, and teachers. States across the country are implementing comprehensive assessment systems. A comprehensive assessment system is "a coordinated and comprehensive system of multiple assessments–each of which is valid and reliable for its specified purpose and for the population with which it will be used–that organizes information about the process and context of young children's learning and development in order to help early childhood educators make informed instructional and programmatic decisions." (US Department of Education definition)

The relevant literature has classified two types of assessment, summative and formative. Summative assessment provides teachers with a snapshot of student understanding. Also called assessment of learning (Stiggins, 2002; Earl, 2005), summative assessments can be a grade on a test or also one on a report card at the end of a marking period. Formative assessment, on the other hand, provides teachers with a tool to ameliorate student achievement while informing instruction (Frobieter, Greenwald, Stecher, & Schwartz, 2011).

Formative assessments are a critical component of comprehensive assessment systems. The definition noted by the Council of Chief State School Officers (CCSSO) seems to best capture the essence of formative assessment. It is defined as, "a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes." (Formative Assessment Advisory Group and Formative Assessment for Teachers and Students (FAST) and The State Collaboratives on Assessment and Student Standards (SCASS), 2006).

Black and Wiliam argue formative assessment is at the heart of effective teaching. In 1998, they conducted a review of more than 250 articles on formative assessment. Based upon the results, they concluded formative assessment does "improve learning." Evidence gathered showed an effect size between .4 and .7. Demonstrating that strengthening formative assessment practices leads to significant and positive learning gains.

The Formative Assessment Process in the Early Childhood Classroom

The process of assessing what young children know and can do poses particular challenges. The traditional approach used for assessing older children is not appropriate for young learners (Ackerman & Coley, 2012; Snow, 2012). In early childhood, each child experiences different rates of growth in their physical, motor, linguistic, and emotional development (Dunphy, 2010; Shepard, et al.). Assessing children is often "unreliable" as young children's performance is not necessarily consistent over even short periods of time, and contextual influences and emotional states are especially relevant for this group (Epstein, Schweinhart, DeBruin-Parecki, and Robin, 2004). In particular, young children develop at vastly different rates and their developmental and learning patterns can be episodic, uneven, and rapid (Bowman, Donovan, & Burns, 2011; Ackerman & Coley, 2012). For these reasons, tests administered at one point in time alone may not provide an accurate picture of the child's concept knowledge, skills, or understanding.

Teachers need an effective evaluation instrument to understand children's development and to help guide their instruction. This instrument should allow them to collect evidence about what students know, determine their skills, and measure their strengths and weaknesses. Reflecting on the data they have collected, teachers can modify their instruction (Büyükkarci, 2014) to identify and reduce gaps in student understanding and provide a pathway for future learning and growth (Black & Wiliam, 1998a; Heritage, 2007, 2008; Sadler, 1989).

An integral part of effective teaching, formative assessment is a systematic process teachers use to gather evidence and provide feedback about student learning, concept understanding, and growth (Black & Wiliam, 1989; Heritage 2007, 2009; Sadler, 1989; Shepard, Kagan, & Wurtz, 1998). By reflecting on student data, teachers determine the current level of understanding, identify gaps in learning, and develop a plan to move toward an educational goal. Most important, teachers use formative assessment to guide their instructional decisions (Stiggins, 2002) when developing plans for, and working with, individual children.

Riley-Ayers, Stevenson-Garcia, Frede, & Brenneman (2012) suggest teachers of young children become participant-observers and engage in an iterative process over time. They can implement a formative assessment process that includes: (1) observing and investigating young children's individual behaviors as a seamless part of instruction, (2) documenting and reflecting on the evidence, (3) analyzing and evaluating the data in relation to set goals or a trajectory of learning, (4) hypothesizing and planning which considers what the children are demonstrating and the implications for instruction, and (5) guiding and instructing where the data helps the teacher target the needs of the children and scaffold their learning to the next level.

The Kindergarten Early Learning Scale

The Kindergarten Early Learning Scale (KELS) was developed in response to a need in the field for a concise observational assessment for young children. Important decisions about the content of the instrument were made, based on several criteria. The items assessed represent the development of kindergarten children, are measurable (observable), develop on a continuum (to see growth and development over time), and are critical to present and future learning (as noted by research).

The KELS examines three domains including (1) Math/Science, (2) Social Emotional/Social Studies, and (3) Language and Literacy, with a total of 10 items across the domains. The items are Number and Numerical Operations, Classification and Algebraic Thinking, Geometry and Measurement, Scientific Inquiry, Responsible Conduct, Habits of Learning, Oral Language, Phonological Awareness, Reading, and Writing. The KELS uses a 5-point continuum with indicator levels at 1, 3, and 5. Scores reported for each of the 10 items are based upon observational evidence collected by the teacher over a period of roughly three months.

Method

Participants

Teachers

In 2013, teachers in 12 counties in a state in the Appalachian region were recruited to participate in the KELS pilot program. A total of 376 teachers participated in the pilot program and 66 teachers from 37 schools from 12 counties were purposely selected to participate in the study based upon the school's location and the teachers' reliability scores. All participating teachers held a bachelor's degree or higher.

Students

As part of the KELS pilot program, each teacher collected anecdotes on 10 randomly selected children in his or her classroom from August 15, 2013 through October 15, 2013. From the 660 children selected by the teachers, NIEER data collectors randomly selected five children to participate in the study from each class. If a child was absent, the next child on the randomly created list was selected. The sample of participating kindergarten children consisted of 276 mostly white children, with a near-split of boys and girls, from 66 classrooms from 37 schools in 12 counties across the state. The mean age of the children was 5.84.

Table 1

		N	%
Gender	Male	139	50.4%
	Female	137	49.6%
Ethnicity	Black	15	5.4%
	Hispanic	4	1.4%
	White	248	89.9%
	Missing	9	3.3%

Training on the KELS

Teachers need to know how to use formative assessment to evaluate a child's progress and also to make positive changes to their teaching pedagogy (Shepherd, et al.). Yet, many teachers enter the field unprepared when it comes to assessment in general, and more specifically assessment for learning (Stiggins, 2002). Additionally, Bergan, Sladeczek, Schwarz, and Smith (1991) question whether teachers (1) know the skills and concepts they should be observing in their students, and (2) know how to observe accurately. Heritage (2007) suggests professional development trainings provide teachers with opportunities to develop (1) domain knowledge, (2) pedagogical content knowledge, (3) knowledge of students' previous learning, and (4) knowledge of assessment.

To help teachers implement the KELS tool successfully in their classroom, tiered approaches were offered for training. Schools and teachers participating in the study committed to a training program, which included both onsite support and online training. A one-day on-site workshop was provided to district coaches working with the school districts in this study. The onsite training program began with an introduction to the KELS, along with a focus on observation and quality documentation. These coaches were used to support the teachers in implementing KELS in the classroom. The amount of support provided varied depending on the coach.

Teachers participated in the online training program (OTP). The program is self-paced and allows teachers to work alone or in groups. Each module in the OTP aligns with best practices for child development, teaching strategies, and current research. This provides teachers with a foundation for using the KELS to inform and improve their instruction and augment student achievement.

Each participant was supplied with a Guide Book. The Guide Book provides detailed information on the KELS, each domain, and each item. Specifically, each item includes a research base, continuum descriptions, ideas for teaching and documenting, sample anecdotes, and a list of resources for further reading. Also included in the Guide Book are the forms needed to implement the KELS including the anecdotal record forms, class record form, and child accomplishments summary form, which is used to communicate with parents regarding the child's development and growth.

The last step in the training process was the teacher's reliability on the instrument. After teachers were trained on the instrument, and implemented the KELS in their classroom for a least one score period for practice and familiarity with the instrument, they were assessed on scoring the KELS. More details about the reliability assessment follow.

Inter-rater Reliability Assessment

Inter-rater reliability was assessed to determine teachers' reliability of scoring data using the KELS instrument. The first step was generating six complete folios for the KELS assessment. These were collected from data in the field and collated to create complete folios with sufficient data to score each of the 10 items. Experts in the field of early childhood education, elementary education, and performance-based assessment, reviewed and scored the folios. An agreed-upon score of 1-5 was determined through discussion and clarification of the evidence for each item in the six folios. The expert score is considered the true score for the item. The teachers were given 3 folios out of the six to review and score using the online system. Agreement with the expert scores determined the teachers' reliability score. The reliability score is a percentage total exact agreement out of 30 items (10 items times three folios). To achieve reliability, teachers had to score 22 of 30 items correctly achieving between 73 and 100 percent agreement.

Concurrent Validity

Standardized, well-established instruments were used to evaluate the concurrent validity of the KELS. These were selected based on use in the field, appropriateness for kindergarten-aged children, and based on the content areas that the KELS examines. The chosen battery consisted of a language (receptive vocabulary), literacy, mathematics, and science assessment. Although the KELS evaluates the social and emotional development domain, a standardized assessment in this domain was not included, because the authors felt that at this time there was not a strong assessment available for the domain that closely matched closely the content of these items on the KELS. Often, social-emotional evaluation comes from teacher self-report measures (e.g., Social Skills Rating System) and this type of reporting would be too closely related to the teacher-reporting of the child's development on the KELS.

NIEER staff trained data collectors (DC) on the three standardized assessment measures (described below). Following the one-day training, data collectors were successfully shadowed by expert staff on two iterations of the assessments for reliability. After two iterations of assessments, each of the data collectors achieved 100% reliability.

Direct Assessment Measures

The *Peabody Picture Vocabulary Test–Third Edition (PPVT-III;* Dunn & Dunn, 1997) is a 204-item test of receptive vocabulary in standard English. The PPVT is predictive of general cognitive abilities and is a direct measure of vocabulary size. The rank order of item difficulties is highly correlated with the frequency with which words are used in spoken and written language. The test is adaptive (to avoid floor and ceiling problems), establishing a floor below which the child is assumed to know all the answers and a ceiling above which the child is assumed to know none of the answers. The test is reliable based on reported split-half reliabilities or test-retest reliabilities. The PPVT has shown concurrent validity (e.g., Qi, Kaiser, Milan, & Hancock, 2006) and the results of these tests are found to be strongly correlated with school success (Blair & Razza, 2007; Early, et al., 2007).

The *Woodcock-Johnson Psycho-Educational Battery-Third Edition (WJ-III;* Woodcock, McGrew, Mather, & Schrank, 2001) includes multiple subtests. Only the *Applied Problems* and *Letter-Word Identification* subtests were used in this study. *WJ* was normed on a stratified random sample of 6,359 English-speaking subjects in the United States. Correlations of the *WJ* with other tests of cognitive ability and achievement are reported to range from .60 to .70. This measure has been used in numerous large-scale preschool studies (e.g., Early, et al., 2007; Wong, Cook, Barnett, & Jung, 2008).

The *Preschool Science Assessment* (*PSA;* Greenfield, Dominguez, Greenberg, Fuccillo, Maier, & Penfield, 2010) is an Item Response Theory (IRT)-based direct assessment of science knowledge and content skills (Greenfield et al., 2012). This assessment was specifically designed and validated to detect growth in children in the Head Start population. The assessment consists of 80 items covering a range of science process skills (e.g., describing, comparing, predicting, experimenting, reflecting) and science content from "life science," "earth and space sciences," and "physical and energy sciences." Children point to the word provided by the test assessor or use manipulatives to display understanding. Pearson reliability was calculated to be

KELS Items

Item 1: Number and Numerical Operations

Item 2: Classification and Algebraic Thinking

Item 3: Geometry and Measurement

Item 4: Scientific Inquiry

Item 5: Responsible Conduct

Item 6: Habits of Learning

Item 7: Oral Language

Item 8: Phonological Awareness

Item 9: Reading

Item 10: Writing

.93 using a Rasch model, indicating a high likelihood that repeated assessment would yield similar scores across children. Discriminant and convergent validity (Osterlind, 2006) were demonstrated. PSA scores improved from fall to spring, showing moderate correlations with math and language scores, smaller positive correlations with approaches to learning, and negative correlations with problem behaviors (Greenfield et al., 2012).

Results

KELS

Table 2 provides KELS descriptive statistics by item and subscale. Oral Language, Item 8, showed highest mean score while Scientific Inquiry, Item 4, showed lowest mean score. The range of scores is 1-5 for each item demonstrating that all items had scores across the full range of possible scores.

Table 2. ***KELS Descriptive Statistics***

	N	Mean	SD	Range
ITEM1	255	2.36	1.02	1 - 5
ITEM2	260	2.31	1.08	1 - 5
ITEM3	241	1.65	1.07	1 - 5
ITEM4	243	1.63	0.86	1 - 5
ITEM5	268	2.43	1.38	1 - 5
ITEM6	267	2.19	1.21	1 - 5
ITEM7	261	2.72	1.25	1 - 5
ITEM8	261	2.56	1.35	1 - 5
ITEM9	259	2.54	1.01	1 - 5
ITEM10	271	2.18	1.02	1 - 5
KELS Math Subscale	266	2.09	0.92	1 - 5
KELS Social/Emotional Subscale	272	2.30	1.21	1 - 5
KELS Language and Literacy Subscale	274	2.48	1.00	1 - 5

Reliability

Reliability concerns the quality of the instrument used. To demonstrate reliability, Creswell (2008) stated, an instrument must be stable and consistent. A reliable research instrument produces clear, consistent, and understandable results (Creswell) in various contexts (McMillan & Schumacher, 2006). McMillan and Schumacher suggested enhancing the reliability of an instrument by offering consistent directions each time it is used, providing the same amount of time to respond to the questions, and conducting the research at the same time of day. The evaluation protocol followed these guidelines and was administered as written.

Internal Consistency

Cronbach's alpha, which demonstrates internal consistency of the measure, was calculated for each of the domains on the KELS. Each domain included between two and four items. All of the three domains showed alpha at .85 - .86. The alpha for the KELS as a whole was .92, indicating that the KELS was measuring a single construct reliably.

Table 3. ***KELS Internal Consistency***

	Cronbach's Alpha
Math	.85
Social/Emotional	.85
Language	.86
Total Items	.92

Inter-rater Reliability.

Inter-rater reliability, also known as criterion-related observer reliability, is the extent to which the trained observer's scores agree with those of an expert observer (Borg, Gall, & Gall, 1989). It is important because it declares that the trained observer understands the variables measured in the instrument with the same efficacy as an expert observer. Table 4 shows that 65 percent of the teachers achieved greater than 60 percent agreement with the true scores on the three reliability folios. The average reliability score was .70.

Table 4. ***KELS Reliability Score Details***

Inter-rater Reliability	Number of Teachers	Percent
.4 to .5	8	12
.5 to .6	15	23
.6 to .7	21	32
.7 to .8	6	9
.8 to .9	2	3
.9 to 1	14	21
Total	66	100

Validity

Validity is a crucial concern when selecting an instrument or instruments for an evaluation study (Lynn, 1986). The *Standards for Educational and Psychological Testing* state, “Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed use of tests” (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 9). A valid instrument such as an observation, interview, questionnaire, or test, should measure what it purports to measure (Kelley, 1927; Lynn; Williams & Monge, 2001). Valid instruments are considered to be accurate and appropriate (Diamond, Luke, & Uttal, 2009; Sullivan, 2011). With certain types of validity, outside subject-matter experts may be asked to weigh in on the validity of the instrument.

Criterion-related validity is how well the test predicts an outcome (Cronbach & Meehl, 1955). One type of criterion-related validity is concurrent validity. Concurrent validity requires both the test and criterion measures be collected at the same time (Creswell, 2008). Using concurrent validity, researchers collect current information about knowledge and skills. This type of validity allows researchers to determine the validity of an instrument by computing a correlation with an existing instrument or instruments (Cronbach & Meehl, 1955). A high degree of correlation between the two instruments provides evidence supporting that the new instrument measures the same underlying dimension equally effectively.

Direct Assessments Descriptive Statistics

Table 5 presents descriptive statistics for the direct child outcome variables in this study for the total sample (N = 276).

Table 5. *Direct Child Assessment Scores*

	N	Mean	SD	Range
PPVT Raw	276	80.81	14.02	33 - 121
WJ1 Letter-Word	276	17.71	6.24	5 - 54
WJ10 Math	275	17.81	3.69	8 - 28
PSA Scaled Score Science	276	615.64	47.94	485 - 767

Concurrent Validity Correlations

Pearson correlation analysis was used to estimate the associations of KELS with direct assessment (see Table 6). The size of these associations varied based on domain. The correlation coefficient between KELS and direct assessment of similar constructs range from .17 to .52. Children’s math and language and literacy scores as assessed by KELS were similar to their scores in direct assessment in those domains: WJ10, PPVT, and WJ1 respectively. The strongest association was found between KELS Language and WJ 1 Letter-Word Identification score (*r* = .52, *p* < .001). Children’s science as assessed by KELS was not as similar to its direct assessment assessed by PSA, though it was significant statistically (*r* = .17, *p* < .01).

Table 6. ***Correlations between KELS Subscale and Direct Child Assessment***

	Math	Science	Social Emotional	Language and Literacy
PPVT Raw	.333***	.286***	.281***	**.428*****
WJ1 Letter-Word	.448***	.313***	.197**	**.519*****
WJ10 Math	**.427*****	.299***	.303***	.453***
PSA Scaled Score Science	.236***	**.167*****	.159***	.389***

To further examine the relationships between the KELS and the direct assessments, correlations at the item level were conducted. Table 7 presents the associations between KELS items and the direct child assessments. Two math items, Item 1 Number and Numerical Operations and Item 3 Geometry and Measurement, demonstrate their highest correlations with WJ10 Math standardized assessment as expected. Item 2, Classification and Algebraic Thinking, significantly correlates with WJ10 math, but has a larger correlation with WJ1 Letter-word. This may be because Item 2 asks children to use language to describe their classifications. Item 7, Oral Language, correlates most highly with the PPVT scores. This makes total sense as the PPVT is a receptive vocabulary assessment closely aligned with language and vocabulary understanding and these skills are tapped into by Item 7 on the ELS. Items 8 and 9, Phonological Awareness and Reading respectively, demonstrate the highest correlations with WJ1 Letter-word which is an assessment of literacy and the expected outcome.

Table 7. ***Correlations between KELS Item and Direct Child Assessment***

	ITEM1	ITEM2	ITEM3	ITEM4	ITEM5	ITEM6	ITEM7	ITEM8	ITEM9	ITEM10
PPVT Raw	.305***	.311***	.273***	.286***	.255***	.272***	**.434*****	.370***	.431***	.222***
WJ1 Letter-word	.405***	.443***	.329***	.313***	.191**	.160**	.420***	**.466*****	**.551*****	.250***
WJ10 Math	**.448*****	**.359*****	**.352*****	.299***	.295***	.265***	.424***	.356***	.449***	.258***
PSA Scaled Score Science	.230***	.257***	.146**	**.167****	.143*	.147*	.326***	.370***	.348***	.197***

Discussion

The mean scores of the KELS assessment for kindergarten children in the first score period of the year (August-October) were consistent with expectations. Means ranged from 1.63 to 2.56 on the five point scale. It is not surprising that the Oral Language item demonstrated the highest mean. This state offers a state-funded prekindergarten-for-all program that enriches children's language experiences during the early years of school. The range of scores on each of the KELS items provides support for the understanding that children enter kindergarten with varying degrees of skills. Using the KELS will provide the necessary evidence for teachers to understand each child's level of development to more accurately plan individualized and intentional instruction.

Findings from this research support the concurrent validity and reliability of the KELS for kindergarten children. The psychometric properties of the ELS are comparable to published instruments in the field of early childhood that use a similar observation approach (Teaching Strategies, 2013; Meisels, Xue, & Shamblott, 2008; Meisels, Liaw, Dorfman, & Nelson, 1995). Teachers were able to achieve acceptable reliability with a mean of .70 on the instrument. This indicates that teachers are able to effectively score data consistently across programs. Further, results demonstrated acceptable levels of validity with moderate relationships with standardized measures in appropriate and meaningful ways. The KELS Oral Language item correlated with the PPVT (.434) and the literacy items on the KELS correlated moderately with the WJ Letter Word Identification Subtest (.446, .551). Similarly for the math domain, the WJ Applied Problems math assessment correlated well with the KELS math items (.448, .359, .352).

With strong support for the relationship of the cognitive components of the KELS with standardized measures we expected to see similar results for the Scientific Inquiry item on the instrument. However, we see that the relationship between this item on the KELS and the PSA has a significant, but rather low correlation. Other published observational instruments similar to the KELS have not reported on the concurrent validity of science items in the literature. This may be because of the lack of effective standardized science assessments for young children. However, with the development of the PSA we decided to include this new instrument in our study. Although the results were not what we expected, several explanations can be offered for the low correlations.

It is important to note that the scientific inquiry process was included in the KELS as an important domain of learning for young children. Specific science content was not included in the instrument, as content topics vary widely from classroom to classroom, and there is not an established consensus of content specific for kindergarten. It is also difficult to place specific content knowledge onto a continuum. However, the scientific inquiry process provides insight across curricula and can be evaluated through any content. This process is a vehicle to learn scientific content knowledge and can easily be assessed in any classroom.

The relationship with the PSA may have been lower as most primary and elementary teachers have not had a solid foundation on how to approach scientific inquiry in their classrooms. Teachers cite a variety of reasons for avoiding science in their classrooms including fear, lack of confidence, dislike for the topic, no pre- or in-service training, or a misunderstanding of science altogether (Appleton, 2003; Davis & Smithey, 2007; Michaels, Shouse, & Schweingruber, 2008; Watters & Diezmann, 1998). Given these reasons for avoiding science, teachers may rely on language and mathematics to teach science. For example, instead of allowing the children to engage in a hands-on science, teachers may conduct an experiment with the children watching and follow it with a discussion. Teachers may ask the children to read or create graphs to document weather patterns or changes in the seasons. Or, teachers may read the science textbook or informational text to the children. These examples provide support for the language and literacy and mathematics domains, rather than for science.

This demonstrates the high use of language and literacy in scientific inquiry and the relationship between math and science in the classroom. Therefore, it seems reasonable that the Scientific Inquiry item on the KELS related more to the PPVT, WJ1, and WJ10 than the PSA (which may be more based in content understanding and scientific background knowledge). Future studies will need to be conducted to further research this relationship. It is our hope that continued work in early science education will yield additional science assessments that will be more closely aligned with the *KELS* approach to science through inquiry.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Appleton, K. (2003). How do beginning primary school teachers cope with science? Toward an understanding of science teaching practice. *Research in Science Education, 33*(1), 1-25. doi:10.1023/A:1023666618800.Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5-25. doi: 10.1080/0969594X.2010.513678.

Black, P., & Wiliam, D. (1998a). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan,* 80(2), 139-148.

Black, P., & Wiliam, D. (1998b). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7–73.

Black, P., & Wiliam, D. (2003). In praise of educational research‘: Formative assessment. *British Educational Research Journal, 29*(5), 623–637.

Black, P., & Wiliam, D. (2004). The formative purpose: Assessment must first promote learning. In M. Wilson (Ed.), Towards coherence between classroom assessment and accountability. Chicago: University of Chicago Press.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation, and Accountability, 21*, 5–31.

Borg, W. R., Gall, M. D., & Gall, J. P. (1989). Educational research: an introduction (5th ed.). New York: Longman.

Büyükkarci, K. (2014). Assessment believes and practices of language teachers in primary education. *International Journal of Instruction, 7*(1), 107-120.

Copple, C., & Bredekamp, S. (Eds.). (2009). *Developmentally appropriate practice in early childhood programs serving children from birth through age 8* (3rd ed.). Washington, DC: National Association for the Education of Young Children.

Cowie, B., & Bell, B. (1999). A model of formative assessment in science education. *Assessment in Education,* 6, 101-116.

Creswell, J. W. (2008). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research.* (3rd ed.). Saddle River, NJ: Pearson.

Cronbach, L. J., & Meehl, P. E. (1955) Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: John Wiley and Sons.

Davis, E. A., & Smithey, J. (2007). Beginning teachers moving toward effective elementary science teaching. *Science Education, 93*(4), 745-770. doi:10.1002/sce.20311.

DeMeester, K., & Jones, F. (2009). Formative assessment for PK–3 mathematics: A review of the literature.

Diamond, J., Luke, J. A., & Uttal, D. H. (2009). *Pratical evaluation guide: Tools for museums and other informal educational settings.* (2nd ed.). Plymouth, UK: AltaMira Press.

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment:

The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research & Evaluation, 14*(7), 1-11.

Dunphy, E. (2010). Assessing early learning through formative assessment: Key issues and considerations. *Irish Educational Studies,* 29(1), 41-56.

Eshach, H., & Fried, N. N. (2005). Should science be taught in early childhood? *Journal of Science Education and Technology, 14*(3), 315–336.

Greenfield, D.B., Dominguez, M.X., Greenberg, A.C., Fuccillo, J.M., Maier, M.F., & Penfield, R. (2010). Lens on Science: *development and initial validation of an Item Response Theory-based assessment of preschool science knowledge and process skills*. Manuscript in preparation.

Heritage, M. (2007). Formative assessment: What do teachers need to know and do? *Phi Delta Kappan, 89*(2) 140-145.

Heritage, M. (2008). Learning progressions: Supporting instruction and formative assessment. Washington, DC: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved

Heritage, M., Kim, J., Vendliniski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative assessment? *Educational Measurement, 28*(3), 24-31.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., et al. (2008). Ready to learn? Children's pre-academic achievements in pre-kindergarten programs. *Early Childhood Research Quarterly, 23*, 27–50.

Keilty, B., LaRocco, D. J., & Casell, F. B. (2009). Early interventionists' reports of authentic assessment methods through focus group research. *Topics in Early Childhood Special*

Education, 28(4), 244–256.

Kelley, T. L. (1927). *Interpretation of educational measurements*. NewYork, NY: Macmillan.

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.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382-386.

Mantzicopoulos, P., Patrick, H., & Samarapungavan, A. (2008). Young children's motivational beliefs about learning science. *Early Childhood Research Quarterly, 23*(3), 378–394.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*(4), 1332-1361. doi:10.3102/0034654309341375

Meisels, S., Liaw, F., 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., Xue, Y., & Shamblott, M. (2008). Assessing, language, literacy, and mathematics skills with Work Sampling for Head Start. Early Education and Development, 19(6), 963–981.

Michaels, S., Shouse, A., & Schweingruber, H. (2008). *Ready, set, science! Putting research to work in k-8 science classrooms*. Washington, DC: The National Academies Press.

McDermott, P. A., Leigh, N. M., & Perry, M. A. (2002). Development and validation of the preschool learning behaviors scale. Psychology in the Schools, 39(4), 353–365.

McMillan, J. H., & Schumacher, S. (2006). *Research in education: Evidenced-based inquiry* (6th ed.). Boston, MA: Allyn & Bacon.

National Association for the Education of Young Children & National Association of Early Childhood Specialists in State Departments of Education. (2003). Early childhood curriculum, assessment, and program evaluation: Building an effective, accountable system in programs for children birth through age 8 [Joint position statement]. Retrieved from http://www.naeyc.org/files/naeyc/file/positions/CAPEexpand.pdf

Ponitz, C. 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), 605–619.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119-144.

Schneider, R. M., Plasman, K., (2011). Science teacher learning progressions: A review of science teachers' pedagogical content knowledge development. *Review of Educational Research, 81*(4), 530-565. doi:10.3102/0034654311423382.

Shepard, L., Kagan, S. L., and Wurtz, E. (1998). Principles and recommendations for early childhood assessments. Washington, DC: National Education Goals Panel. Retrieved www.negp.gov/reports/prinrec.pdf

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

Sullivan, G. M. (2011). A primer on the validity of assessment instruments. *Journal of Graduate Medical Education, 3*(2), 119-120. doi:10.4300/JGME-D-11-00075.1

Teaching Strategies (2013). Teaching Strategies Gold Assessment System: Concurrent Validity. Retrieved http://teachingstrategies.com/content/pageDocs/GOLD-Concurrent-Validity-2013.pdf

Watters, J. J., & Diezmann, C. M. (1998). "This is nothing like school": The constructivist learning environment for early childhood science. *Early Childhood Development and Care, 140,* 73-84. doi:10.1080/0300443981400106.

William, F., & Monge, P. (2001). *Reasoning with statistics: How to read quantitative research.* (5th ed.). Belmont, CA: Thomson Higher Education.

Woodcock, R. W., McGrew, K. S., Mather, N., & Schrank, F. A. (2001). *Woodcock–Johnson III (WJ-III)*. Itasca, IL: Riverside.