Journal of Advanced Academics
jaa
Volume 20 Number 3 Spring 2009 pp. 502–535

The Development and Validation of the Science Learning Assessment (SLA):

A Measure of Kindergarten Science Learning

Ala Samarapungavan
Panayota Mantzicopoulos
Helen Patrick
Brian French
Purdue University

Introduction

Over the last few decades there has been a growing call to develop reform-oriented and inquiry-based science instruction (American Association for the Advancement of Science [AAAS], 1993; National Research Council, 1996). Recently a few programs such as the Head Start on Science and Communication Program (HSSC; Klein, Hammrich, Bloom, & Ragins, 2000), the ScienceStart! Program (French, 2004), and the Preschool Pathways to Science (PrePS; Gelman & Brenneman, 2004) have begun to develop innovative preschool and kindergarten science curricula. However, there is a lack of instructionally sensitive assessments that can be used to measure, aggregate, and compare

Summary

The Science Learning Assessment (SLA) is an individually administered, instructionally sensitive science assessment for kindergarten students. The SLA is a 24-item objective test, broken down into two subtests. The Scientific Inquiry Processes subtest consists of 9 items designed to measure young children's functional understanding of the nature and processes of scientific inquiry. The Life Science Concepts subtest consists of 15 items designed to measure children's understanding of specific science concepts related to living things and the physical world. Our results on SLA items that assess life science concepts indicate that kindergarten children are able to develop a rich knowledge base about living things. The results of the SLA indicate that even young children can begin to develop an understanding of scientific inquiry with appropriate instructional support. Our findings are consistent with recent work by researchers such as Metz (2004), who argue for a richer conception of children's developmental capacities in the context of science instruction. As educators develop new science curricula and programs to address the lack of rich and challenging science instruction in the early grades, there is a need to document what children learn from such efforts. In order to develop research-based and pedagogically effective science curricula, we need assessments with clearly described theoretical and psychometric characteristics. The SLA is one example of such an assessment that can be used to aggregate and compare learning outcomes, as well as to provide empirical information on kindergarten children's capacities for science learning.

Samarapungavan, A., Mantzicopoulos, P., Patrick, H., & French, B. (2009). The development and validation of the Science Learning Assessment (SLA): A measure of kindergarten science learning. *Journal of Advanced Academics, 20,* 502–535.

learning outcomes for young children, especially across different instructional approaches. The purpose of this paper is to describe the development and initial validation of such an instrument, the Science Learning Assessment (SLA).

The SLA was developed in the context of the Scientific Literacy Project (SLP; Mantzicopoulos, Patrick, & Samarapungavan, 2005) to provide a measure of kindergarten students' science learning. It is designed as a proximal assessment in a system of multilevel assessments of learning. Ruiz-Primo, Shavelson, Hamilton, and Klein (2002) characterized as "proximal" assessments that measure children's knowledge of the specific concepts taught in the curriculum, but may include items or tasks that are different from the topics in the context of which those concepts are taught in the classroom. For example, while children might learn about the processes of scientific inquiry through a unit on the growth and development of butterflies, a proximal assessment might assess their understanding of inquiry processes through items that include other or different content.

The SLA is a test of kindergarten students' science knowledge. Specifically, it is designed to assess the science learning of students who receive an inquiry-based science curriculum as part of the SLP project. The science activities that comprise the kindergarten science curriculum are developed collaboratively by SLP project personnel and teachers in the participating schools and implemented by the teachers. Each set of science activities is mapped to the Indiana academic science standards for kindergarten (Samarapungavan, Mantzicopoulos, & Patrick, 2008).

The development of the SLA is guided by the view of science learning as a process of domain-specific knowledge construction (Carey & Spelke, 1994; Gelman & Brenneman, 2004) that is socially negotiated and situated in specific cultural contexts and practices (Boyd & Richerson, 2005; Brown, 1997; Driver, Asoko, Leach, Mortimer, & Scott, 1994; Greeno, 1998; Rogoff, 1990). Consistent with this perspective, the SLA is designed to be instructionally sensitive to measure specific science learning outcomes that are fostered in the kindergarten science curriculum implemented in SLP classrooms. Before

describing the development of the SLA in more detail, it is useful to consider how other researchers have approached the problem of assessing the impact of specific science curricula on early science learning.

Assessments of Learning in Research on Kindergarten Science Programs

Embedded Assessments

The term *embedded assessments* refers to measures whose assessment format is embedded in, and therefore strongly tied to, the specific instructional practices of classrooms. Examples include various kinds of performance assessments such as portfolio ratings. Ruiz-Primo et al. (2002) referred to such assessments as "immediate." The benefits of such assessments have been widely discussed. It has been proposed that they have consequential validity (Messick, 1994) because they provide students the opportunity to learn as they are being assessed, allow teachers to more accurately diagnose what students know by revealing children's thinking, help teachers adapt instruction to students' thinking, and can document students' cognitive growth over time (Gitomer & Duschl, 1998; Sheppard, 2000). Researchers have shown that performance assessments of young children can be designed to have adequate psychometric properties of reliability and validity (Meisels, Liauw, Dorfman, & Nelson, 1995). The current (SLP) project also utilizes a variety of embedded assessments such as portfolio ratings, evaluations of science notebooks, and analysis of child discourse to measure science learning (Samarapungavan et al., 2008). However, one limitation of such assessments is that they do not provide any direct basis for comparison across instructional programs. For instance, kindergarten programs that do not use inquiry-based pedagogies typically do not generate similar artifacts or performances for comparison.

Standardized and Norm-Referenced Assessments

In the absence of developmentally appropriate measures of science learning, some projects such as ScienceStart! have used general standardized measures of cognitive growth such as the Peabody Picture Vocabulary Test (PPVT) to compare student gains across instructional programs (French, 2004). The PPVT and other similar measures provide an index of skills (vocabulary and cognitive skills) that may be indirectly enhanced by systematic efforts to teach children science. However, these assessments tell us little about the nature and extent of children's science learning across programs.

One science-specific standardized measure of young children's science achievement is the Science subtest of the Woodcock Johnson III (WJ-III) battery (Woodcock, McGrew, & Mather, 2001). This individually administered measure is one of three subtests (social studies, humanities, and science) that comprise the Academic Knowledge cluster of the WJ-III. The Science subtest is designed to assess general knowledge in biological and physical sciences and includes items appropriate for preschool children. For the early items, questions are accompanied by pictures, one of which represents the correct response. The child is asked to indicate his or her response by pointing, and it is argued that these items are suitable for young children who may not yet possess sufficient verbal skills to express their knowledge. However, after the first seven items (the recommended starting point for kindergarten children), the response format changes to one that demands an oral response to the examiner's question.

Normative data are only provided for the entire Academic Knowledge cluster and are not specific to the Science subtest. Moreover, items were selected to provide a "broad sampling of achievement" (Schrank, McGrew, & Woodcock, 2001, p. 15) and do not assess children's domain-specific knowledge and skills across key science themes and concepts. Data on the validity of this measure as a global indicator of general science knowledge are not available and there is no evidence of the test's sensitivity to instructional or schooling effects over time.

The items of the WJ-III Science subtest draw on a narrow set of vocabulary and general knowledge skills and prompt for recall of labels for things (e.g., names of animals) or processes. One question asks children to provide the word that describes the process of littering (pollution). Unlike the SLA, the WJ-III Science subscale for kindergarten does not include items that probe young children's conceptual understanding of scientific inquiry processes such as observing, predicting, measuring and recording data, or hypothesis testing. In this study, we use the WJ-III Science subtest as a rough indicator of children's general knowledge about science.

Unit Tests of Conceptual Knowledge

Some programs such as ScienceStart! and HSSC assess knowledge gained from specific science units in their curriculum with researcher-designed questions that are administered before and after the unit is completed (French, 2004; Klein et al., 2000). Assessments of the kind described here are examples of "close" assessments. Although the exact nature and content of such questions are not described in detail in the literature, it appears that these assessments require a fair degree of receptive as well as expressive language proficiency from the child. For example, questions in the HSSC program ask children to verbally explain what they learned (Klein et al., 2000). ScienceStart! employs a narrative assessment format in which students read storybooks about a protagonist called Curi the curious bear. At various points in the story, children are presented with a problem encountered by Curi and must explain what Curi could do to solve the problem before they can continue on with the story (French, 2004). The published literature on such approaches to assessment does not describe how such tests are administered and how children's verbal responses are scored. There is no information about the psychometric properties of the assessments (their internal consistency or validation procedures and evidence). It also appears that these assessments measure children's understanding of specific science content, such as color, but do not measure their understanding of scientific inquiry.

The current research on the SLA synthesizes and extends recent efforts to develop science assessments that allow us to aggregate and compare science learning across instructional programs. The process of test development and validation was informed by the guidelines outlined in the National Educational Goals Panel Report titled, "Recommendations and Principles for Early Childhood Assessments" (Shepard, Kagan, & Wurts, 1998). An overview of the process by which the SLA was developed follows.

Developing the SLA: An Overview

Conceptual Content of the SLA

As noted earlier, the development of the SLA is guided by certain theoretical assumptions about what it means to learn and to know science. Because science learning is viewed as a process of domain-specific knowledge construction, the first step in developing a test blueprint was to map out the key aspects of science targeted in the SLP project curriculum in Year 1 (see Table 1).

The curricular goals for SLP and assessment objectives for SLA were synthesized from the content standards for kindergarten science learning specified in three standards documents: the National Science Education Standards (Center for Science and Mathematics Education, 1996), the Benchmarks for Scientific Literacy (AAAS, 1993), and Indiana's Academic Standards: Kindergarten Science (Indiana Department of Education, 2006).

The SLA items designed to assess target concepts specified in Table 1 were developed and reviewed by content area experts in science and experts in early science learning who provided support for the content validity of this assessment. This is consistent with the recommendations of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Descriptions of SLA items are provided in Tables 2 and 3. Additional details

Table 1

Mapping of SLA Items to National and State (Indiana) Science Education Content Standards for Kindergarten

Target Understanding	*SLA Items**
Scientific Inquiry Processes	
• Understand science as a process of inquiry is based on asking questions and making predictions about the natural world	1, 2, 3, 4, 5
• Understand the empirical basis of science: Scientific ideas are evaluated by their correspondence or fit to empirical evidence	1, 2, 3, 4, 5, 6, 7, 8, 9
• Understand simple tools used to gather, record, analyze, and share data	19, 20, 21
Life Science	
• Understand the characteristics of living things. For example, they need air, water, and food; they respond to their environment; they reproduce etc.	10, 11, 12, 13, 14, 15, 16, 18, 22, 23, 24
• Structure and function: Understand that plants and animals have specific structures and traits (e.g., physical and behavioral characteristics) that help them survive, grow, and reproduce	
• Understand that living things have life cycles: They are born, develop into adults, reproduce, and eventually die	17

Note. **Descriptions of SLA items are provided in Tables 2 and 3.**

about item format, test administration, and scoring are provided in the Methods section below.

Methods

Description of SLA

SLA Materials. The SLA contains 24 items: Nine items assess children's understanding of scientific inquiry processes (see Table 2) and 15 items assess their understanding of life science concepts (see Table 3).

Item format. Most SLA items follow a format in which the child is shown three pictures (each on a separate card) and

Table 2

Description of SLA Items to Test Understanding of Scientific Inquiry Processes

1. Here are picture of three children (show pictures): Which of these children is doing science? (a) *Gina observes a butterfly*; (b) Tom plays the guitar; (c) James practices dancing.

2. Here is a picture of a frog (show picture). These girls ask questions about the frog (show pictures). Listen to each question and tell me which girl asked a science question: (a) *What does this frog eat*? (b) Do you like this frog? (c) Can I call this frog Lilly?

3. Here is a picture of a fish (show picture of black and white striped fish). Here are three boys (show pictures). I will tell you what each boy said about a fish. (a) *That fish has black and white stripes*; (b) I have a pet goldfish at home; (c) Fish like to swim in groups. Which of these boys saw the fish in this picture?

4. Here is a picture of a ball (show picture of red ball at rest). Here are three girls (show pictures). I will tell you what each girl said: (a) *This ball can bounce*; (b) This ball is red; (c) My dress is green. Which of these girls made a prediction about the ball?

5. Tony, John, and Gina are on the playground (show picture: John and Gina are on the teeter totter. John is on the end that is down to the ground and Gina is on the side that is up in the air. Tony is standing beside them.) Listen to what each child says. Then tell me which child makes a prediction about the teeter totter: (a) *Tony says, "If I push down on Gina's side, John will go up;"* (b) John says, "I want to go up Gina;" (c) Gina says, "I am having lots of fun." Which of these children makes a prediction about the teeter totter?

6. (Show pictures) Two girls found an egg. The girl in green thinks it is a duck egg. The girl in blue thinks it is a goose egg. How can they find out what it is? (Correct answer: Watch it hatch, study its shape, etc.)

7. Here are some tools we use to do science (show pictures): Which of these can you use to help you remember what you saw? (a) *Science notebook*; (b) Magnifying glass; (c) Stopwatch

8. Here are some tools we use to do science (show pictures): Which of these can you use to look at something very small such as a bug? (a) *Microscope*; (b) Rain gauge; (c) Digital scale

9. Here are some tools we use to do science (show pictures): Which of these can you use to measure how hot something is? (a) *Thermometer*; (b) Rain gauge; (c) Pan scales

***Note*. Pictures are presented in random order within items.**

Table 3

Description of SLA Items Testing Understanding of Life Science Concepts

1. One of these animals is an insect (show pictures): Which one? (a) *Butterfly;* (b) Bird; (c) Squirrel
2. (Follow up to 1). How do you know that this is an insect? (Correct answer: Six legs)
3. One of these animals is an insect (show pictures): Which one? (a) *Ant;* (b) Centipede; (c) Spider
4. (Follow up to 3). How do you know that this is an insect? (Correct answer: Six legs, 3 body segments)
5. This is a picture of a caterpillar (show picture). What parts of the body does the caterpillar use to eat? (Correct answer: Says "mouth" or points to mouth on picture)
6. (Use picture for 5). What parts of the body does the caterpillar use to breathe? (Correct answer: Says "holes on side" etc., and/or points to spiracles)
7. What parts of the body does the caterpillar use to move? (Correct answer: says "legs" and/or points to legs)
8. These pictures show how a butterfly is born and grows and changes through its life (show pictures in following order: egg; monarch caterpillar; empty slot; monarch butterfly). Look at the three pictures below (show pictures): (a) *Monarch chrysalis*; (b) House fly; (c) Miniature version of monarch butterfly. Which of these should go up here (point to missing picture) to complete the butterfly life cycle?
9. Here are some pictures of plants (show pictures of two plants, one with green foliage and one with orange and brown foliage). On which plant should this butterfly stay (show picture of monarch butterfly) so that it won't be seen? (Correct answer: plant with orange and brown foliage)
10. Here are some pictures of animals (show pictures). Which of these in NOT camouflaged? (a) *Orange goldfish in green pond water*; (b) Brown/grey toad on brown/grey tree trunk; (c) Grey moth on grey tree bark
11. Which of these is a living thing (show pictures)? (a) *Plant*; (b) Car; (c) Table
12. (Follow up to 11). Why is it a living thing? How can we tell that it is a living thing? (Correct answer: Names two or more characteristics of living things; e.g., grows, needs food or water, breathes, moves on its own, etc.)
13. (Show pictures) One of these needs air to breathe: Which one? (a) *Dog*; (b) Doll; (c) Balloon
14. (Show pictures) One of these needs food: Which one? (a) *Ant*; (b) Robot; (c) Bicycle
15. (Show pictures) One of these has no back bone: Which one? (a) *Crab*; (b) Damselfish; (c) Girl

***Note*. Pictures are presented in random order within items.**

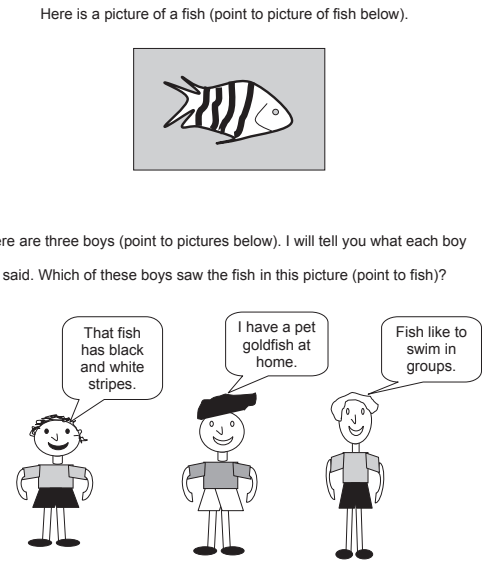

***Figure 1*. Schematic representation of SLA Item 3 with grayscale versions of actual color pictures.**

asked a question about these pictures that can be answered verbally or by pointing to the correct pictures (see Figure 1 for an example). If the child's response is not clear, the examiner follows up by asking the child to repeat his or her answer, to point to the picture again, or to explain his or her answer (e.g., I am sorry, I did not hear you the first time. Could you say that one more time? Could you tell me a little more? Could you show me which picture you chose?). If the child does not respond, the examiner repeats the entire question sequence once. If there is still no response, the examiner moves on to the next question. Although the correct answer for choice format items is always shown first in Tables 2 and 3 and recorded as option "a" on the scoring response sheet (see Figure 2), the actual order in which the response choice cards is presented is randomized for each child. The order in which the response choices are presented is also recorded on each child's response sheet (see Figure 2).

In addition to the choice format described above, the SLA also includes some open-ended follow-up questions. An exam-

SLA	Child's Name:		ID#	Test Date:	Teacher:	
Please Present Response Choices in Random Order to Each Child and Record Order in 2nd Column Below						
Question	Order	Correct Response	Incorrect Responses or No Response			Score
1. Doing science		a. Gina: Observes butterfly	b. Tom: Plays guitar	c. James: Practices dancing	NR	a = 1 b /c / NR = 0
2. Science question		a. What does this frog eat?	b. Can I call this frog Lilly?	c. Do you like this frog?	NR	a = 1 b / c / NR = 0
3. Observed fish		a. Fish has black and white stripes	b. Have goldfish at home	c. Fish like to swim in groups	NR	a = 1 b / c / NR = 0

Figure 2. **Excerpt from SLA response sheet.**

ple is SLA Item 4 (see Table 3), which is the follow up to Item 3 (insect identification).

General Administration Procedures. The SLA is individually administered. The examiner and child are seated at a table, at right angles to each other. The examiner has a test packet that includes: the manipulatives for each item in individually labeled envelopes, a printed booklet with the examiner's script and pictures for each item (kept out of the child's view), and a sheet to record the child's response or choice for each item (see Figure 2). The examiner begins the testing session with the following introduction:

> *Hello, (say child's first name). My name is _____. I am from _____. Thank you for helping us out today. We would like to know what children like you think about science and the world around you. This is not a test for school. We just want to know what you think. You do not have to go on if you do not want to. You can stop at any time. Would you like to go on? (If child gives assent) I will ask you some questions about science. There are pictures that go with each question. Let us start with the first question.*

The manipulatives for SLA Item 1 are then laid out on the table and the examiner follows the script for that item. After the child responds, the examiner records the response on the response sheet and returns the manipulatives to their envelope. The examiner then proceeds to administer the next item until all items have been administered. If the child wishes, he or she is allowed to return each set of manipulatives to its envelope. The test is terminated if the child fails to respond to five consecutive items.

Scoring of the SLA. A dual coding scheme is used to score SLA responses. Correct responses are scored 1 while incorrect responses and nonresponses are scored 0 (see Figure 2 for example). Thus, the possible SLA-Total scores range from 0 (*none correct*) to 24 (*all correct*). Subscores are computed by summing

scores for the SLA-Scientific Inquiry Processes items (Items 1–9) and the SLA-Life Science Concepts items (Items 10–24), respectively.

Participants

This study was conducted in a Midwestern, suburban public school district. Data were collected from 100 kindergarten children in two different schools who participated in the Scientific Literacy Project. In School 1, there were 82 children in four kindergarten classrooms who were enrolled at the beginning of the science unit. We obtained informed consent for 71 (86.5%) of those children. However, we collected SLA data for 65 children because 6 children moved from the school before the end of the unit and could not be tested. The 65 children in School 1 comprised our intervention group. They participated in a 5-week-long SLP science inquiry unit prior to the administration of the SLA.

In School 2, there were 52 children in two kindergarten classrooms, and we obtained informed consent for 35 (68%) of them prior to SLA administration. SLA data were collected for all 35 children. These 35 children comprised our comparison group and did not participate in any SLP project activities prior to being tested with the SLA. To control for general instructional effects, both the intervention and comparison group were assessed in the same 2-month period around the middle of the academic year. Chi-square tests were conducted to examine the comparability of the children in the two groups on gender, ethnicity, and free or reduced-cost lunch status. There were no statistically significant differences on these variables.

Ethnicity information for the sample was as follows: Fifty-nine percent of the sample was Caucasian, 11% was African American, 20% was Hispanic, and 10% was classified as other. There were 56 boys and 44 girls. Free lunch information was available for 81 of the children, 56 (69.1%) of whom received free or reduced-cost lunch.

Description of Instructional Context

SLP Activities in School 1. The theoretical framework for the SLP project centers on the notion that young children develop scientific literacy in everyday interactive contexts through cognitively guided learning and discourse about science with adults (teachers and parents). Classroom activities are built around science topics and skills linked to the state academic standards for kindergarten and are mapped to science, English/language arts, and mathematics standards. They also are consistent with current guidelines for developmentally appropriate practice (National Association for the Education of Young Children, 2003).

In School 1, the kindergarten teachers and their students engaged in a sequence of inquiry activities to explore the properties of living things and the theme of growth and development through the life cycle of the monarch butterfly. The inquiry unit, which was developed for SLP, integrated a variety of inquiry and literacy activities and was implemented twice a week over a 5-week period (see Samarapungavan et al., 2008, for a detailed description of the intervention). As part of the unit, children engaged in learning activities grouped into three broad phases as follows:

1. *Pre-inquiry activities* introduced children to key ideas about the nature of science, provided them with the procedural framework for their investigations, and incorporated book readings to provide relevant background knowledge.
2. *Inquiry activities* were centered on the growth and development of live monarch larvae on milkweed plants. Children started by formulating questions and making predictions about the larvae's growth and development, then they observed and recorded what happened in their science notebooks (using words, digital images, and drawings), and finally they drew conclusions about the monarch life cycle based on their records. During this phase, children read books about the characteristics of insects to extend their knowledge.

3. *Post-inquiry activities* were designed to help children reflect upon and communicate what they had learned (e.g., through posters that they shared). During this phase children read books on the butterfly life cycle to systematize what they had learned through their own inquiry.

Science Activities in School 2. The kindergarten teachers in School 2 said that they did not "teach science" because of the heavy focus on literacy and numeracy, although they sometimes read books on topics such as animals and seasons as part of their literacy activities. They did report that their classes had taken an informal field trip to a nearby nature conservancy.

External Measures Used to Provide Evidence of SLA Validity

This study used data from two independent measures, gathered as part of the SLP project, to examine the validity of the SLA. The measures were: (a) the Science Knowledge and Passage Comprehension subtests from the Woodcock-Johnson III Tests of Achievement III (WJ-III; Woodcock et al., 2001), and (b) the SLP Portfolio Rubric.

WJ-III Science Knowledge and Passage Comprehension Subtests. The Science Knowledge subtest was described in the introduction and is one of several subtests that comprise the WJ-III Tests of Achievement. The Science Knowledge subtest is designed to measure general biological and physical science knowledge. Psychometric information is available only for the full Academic Knowledge cluster. Reliabilities are .84 (1 year test-retest reliability for 2- to 7-year-olds) and .92 (split-half-reliability for 4- to 6-year-olds). To gauge the reliability of scores in our sample on the Science Knowledge subscale, we computed Cronbach's alpha with the children participating in this study. The alpha coefficient was .67. It is of note that this estimate is lower than the alpha coefficients for the SLA full scale as well as each of its subscales (see next page).

The Passage Comprehension subtest is part of the WJ-III Broad Reading cluster, and it provides information on the child's vocabulary and comprehension skills and ability to understand language when it is being read. This subtest requires use of semantic and syntax cues as the child identifies missing information in each question. The test-retest reliability reported over a 1- to 2-year interval for this subtest is .75 for children in the 2- to 7-year-old age range (McGrew & Woodcock, 2001). Internal consistency reliabilities range from .94 to .96 for children in the 4- to 7-year-old range. Correlations of the WJ-III Broad Reading cluster with commonly used achievement measures range from .25 to .76 across subtests of the Kaufman Test of Educational Achievement and the Wechsler Individual Achievement Test. Ruiz-Primo et al. (2002) characterized achievement tests such as the WJ-III as "distal" measures in a multitier system of assessment.

The SLP Portfolio Rubric. SLP employs an electronic portfolio system to sample and evaluate children's work on each inquiry unit. The children's portfolios contain electronic records of artifacts (science notebook entries) that they produced, excerpts from (videotaped and transcribed) classroom discourse, and written comments by the classroom teacher or teacher's assistant. SLP personnel are trained to use the portfolio system to document and evaluate children's learning from SLP classroom activities. The artifacts and records that comprised the SLP portfolio were selected by the researchers in collaboration with the classroom teachers. For a detailed description of the SLP Portfolio Assessment, see Samarapungavan et al. (2008).

Portfolio evidence was collected and evaluated for the 65 children in School 1 who participated in the inquiry unit on the life cycle of the monarch butterfly. Children's portfolios were rated by trained raters on seven specific aspects of knowledge and skill (Portfolio Items P1–P7). Interrater reliability for the portfolio ratings was $r = .91$; $p < .01$ (Samarapungavan et al., 2008). Items P1–P4 represent understanding of the processes of scientific inquiry such as the ability to ask questions about

the natural world, to gather and use empirical evidence, and to communicate about science. An example of these items is: "P1. Raises questions / makes predictions about the natural world." Items P5–P7 reflect children's understanding of life science concepts, especially with regard to growth and development. An example of these items is: "P7. Understands the growth and development of the monarch butterfly." Scores on individual portfolio items ranged from 0 to 3 (higher scores indicate greater mastery) and a Composite Portfolio Score was obtained for each child by adding the individual scores across the seven items in the portfolio rubric. Portfolio subscores were computed as follows: (a) The Portfolio Inquiry Processes score was computed by summing item scores across items P1–P5, and (b) The Portfolio Life Science Concepts score was computed by summing scores across items P6–P7.

The portfolio rubric measure is an example of what Ruiz-Primo et al. (2002) called immediate assessments, those based on and drawn directly from a classroom learning activity. Thus, as part of a system of multitier assessments of learning we can examine the extent to which measures at different levels, such as the WJ-II Science Knowledge subtest and the SLP Portfolio Assessment, correlate with the SLA.

Summary of Assessment Procedures

At the conclusion of the SLP curriculum activities, the children in School 1 were tested with the SLA and WJ-III subtests. Testing took place in two sessions to keep the administration time short and maintain children's interest and engagement with the tasks. Children in School 2 were tested with the SLA within the same 2-month period. In School 1, after the inquiry unit was completed and all unit artifacts and video data were entered into the electronic database system, a member of the SLP project scored the electronic portfolios.

Analytic Plan

Several analyses of the SLA data were conducted to examine the psychometric properties of the SLA and provide validity evidence. These are outlined below.

Internal Consistency

To examine the internal consistency of the SLA, we computed internal consistency reliability coefficients (Cronbach's alpha) for scores for the two components of the SLA (SLA-Scientific Inquiry Processes and SLA-Life Science Concepts) as well as for the full-scale score.

Consistency With Theoretical Constructs

Confirmatory Factor Analysis. To further explore whether the pattern of responses was consistent with the theoretical goals of measuring two aspects of science knowledge (understanding of scientific inquiry processes and understanding of life science concepts respectively), we conducted a confirmatory factor analysis to examine whether the responses fit a two-factor model.

Instructional Sensitivity. To examine whether the SLA was instructionally sensitive, we used multivariate analysis of variance (MANOVA) to analyze performance differences between the intervention and comparison groups (Schools 1 and 2, respectively).

Item Difficulty. We examined the means and standard deviations for each SLA item to determine whether the test represented a range of difficulty levels and would be sensitive to differences in children's science knowledge. We also reviewed these results to determine whether the observed patterns of item difficulty are consistent with theoretical expectations and prior research on young children's science learning.

Correlations With Other Measures of Achievement and Science Learning. To examine how the SLA corresponds with other measures of achievement and learning, we computed bivariate correlation coefficients (Pearson *r*) between SLA scores and each of the following measures: (a) WJ-III Science Knowledge subtest (raw scores), (b) WJ-III Passage Comprehension subtests (raw scores), and (c) SLP Portfolio Assessment scores. Portfolio Assessment scores were available only for the 65 children in the intervention group who participated in SLP activity as no portfolios were maintained in the comparison school.

Results

Internal Consistency

The reliability analyses indicate that our sample's scores on the SLA exhibit adequate internal consistency given its purpose. The alpha coefficient for the data obtained from 9 items that assessed children's understanding of Scientific Inquiry Processes was .71. The alpha coefficient for the data from the 15 items that assessed children's knowledge of Life Science Concepts was .70. The full-scale alpha for data from all 24 items in the SLA was .79.

Consistency With Theoretical Constructs

Confirmatory Factor Analysis. The item-level CFA was conducted with M*plus* 3.11 (Muthén & Muthén, 2004). A robust weighted least squares (i.e., WLSMV) estimation was used to account for the dichotomous data, per the recommendation of Finney and DiStefano (2006). WLSMV does result in greater stability of estimates with small samples and with various factor models, number of variables, and number of response options compared to standard WLS estimation (Beauducel & Herzberg, 2006; Flora & Curran, 2004). Because the testing of rival models provides stronger validity evidence (Thompson & Daniel, 1996), two competing models were tested: the theoretically preferred

two factor model (Model A), and one in which the two factors were simply facets of a single factor construct (Model B).

Model fit was evaluated by the chi-square significance test, comparative fit index (CFI), Tucker-Lewis fit index (TLI), and the Weighted Root Mean Square Residual (WRMR). Note that the degrees of freedom for the chi-square test with WLSMV estimation is estimated and not computed in standard fashion (see M*plus* user's guide). WRMR applies only to WLSMV estimation; values less than 1.0 indicate good fit (Yu & Muthén, 2002). CFI and TLI both indicate the relative fit of a given model as compared to a null model. Values above 0.90 may be an indication of adequate fit. The use of multiple fit criteria in combination follows recommendations to examine combinations of fit indices (Hu & Bentler, 1999).

The fit of Model A was better than that of Model B ($\chi^2(39) = 93.12, p < 0.05$, CFI = 0.79, TLI = 0.80, WRMR = 1.31; $\chi^2(39) = 107.45$, $p < 0.05$, CFI = 0.73, TLI = 0.74, WRMR = 1.41), respectively). In reviewing item loadings, loadings for three items (Items 6, 14, and 16) were quite low (0.17, 0.24, and 0.16, respectively). The corrected item-total correlations for these items with their respective subscales as well as with the SLA whole scale were as follows: Item 6 had a corrected item-total correlation of 0.22 with the Scientific Inquiry Processes subscale and 0.10 with the SLA whole scale. Item 14 had a corrected item-total correlation of 0.25 with the Life Science Concepts subscale and 0.15 with the SLA whole scale. Item 16 had a corrected item-total correlation of 0.39 with the Life Science Concepts subscale and 0.07 with the SLA whole scale. Upon the removal of these items, fit for the two factor model improved ($\chi^2(37) = 70.03, p < 0.05$, CFI = 0.89, TLI = 0.89, WRMR = 1.17). Note that total scale score internal consistency reliability only changed from 0.79 to 0.80. The subscale alpha for Scientific Inquiry Processes after deleting Item 6 changed from 0.71 to 0.72 while the subscale alpha for Life Science Concepts after deleting Item 14 and Item 16 was unchanged at 0.70.

Table 4 provides the parameter estimates for Model A with the original 24-item scale as well as with items 6, 14, and 16 from

Table 4

Standardized Pattern Coefficients and Uniqueness Estimates for the Two-Factor Model

Estimates for Original 24-Item Scale

Understanding of Scientific Inquiry Processes			*Understanding of Life Science Concepts*		
Item	Pattern	Uniqueness	Item	Pattern	Uniqueness
1	0.797	0.354	10	0.344	0.882
2	0.492	0.757	11	0.805	0.343
3	0.946	0.105	12	0.841	0.292
4	0.939	0.118	13	0.823	0.323
5	0.465	0.784	14	0.247	0.939
6	0.165	0.973	15	0.770	0.407
7	0.672	0.548	16	0.159	0.975
8	0.629	0.605	17	0.376	0.859
9	0.223	0.950	18	0.727	0.471
			19	0.350	0.878
			20	0.725	0.475
			21	0.603	0.636
			22	0.460	0.788
			23	0.483	0.767
			24	0.377	0.858

Estimates for Revised 2-Item Scale (Items 6, 14, and 16 deleted)

Understanding of Scientific Inquiry Processes			*Understanding of Life Science Concepts*		
Item	Pattern	Uniqueness	Item	Pattern	Uniqueness
1	0.791	0.374	10	0.300	0.910
2	0.451	0.797	11	0.803	0.355
3	0.949	0.100	12	0.833	0.307
4	0.933	0.130	13	0.812	0.340
5	0.451	0.796	15	0.774	0.402
7	0.666	0.556	17	0.387	0.850
8	0.614	0.623	18	0.756	0.428
9	0.260	0.932	19	0.361	0.870
			20	0.704	0.504
			21	0.583	0.660
			22	0.464	0.784
			23	0.494	0.756
			24	0.384	0.853

the original SLA removed. The correlation between the factors in the 24-item and the revised 21-item scale were 0.61 and 0.65, respectively. We decided to retain Model A because the factor analysis indicated a better fit for this model over the alternative, Model B, even though it did not meet strict model fit guidelines (e.g., CFI > 0.90), because the literature indicates that it is not clear how these guidelines function with RWLS (Beauducel & Herzberg, 2006), and no other model modifications were empirically indicated or theoretically justified. As seen in Table 4, each of the items on their respective factors had generally moderate to high pattern coefficients and were quite variable (range of loadings = 0.27 to 0.951). Although model fit was improved by deleting Items 6, 14, and 16, these items represent important content from a developmental and a science learning perspective. Item 6 is the only SLA item that measures children's understanding of the concept of hypothesis testing, which is an important component of the processes of scientific inquiry. Although, as we elaborate in the discussion below, the developmental literature suggests that this is a difficult concept for young children, it is possible more sustained and longer lasting inquiry-based instruction could help kindergarten students learn this concept, which would be reflected in a higher proportions of correct response on this item *and* which in turn would improve its factor loadings and corrected item-total correlations. This is also an item that could potentially differentiate between high- and low-achieving students because of its difficulty level. Items 14 and 16 also test important biological concepts in the SLP curriculum as they both assess children's understanding of the relationship of biological structure and function, using the monarch caterpillar as an exemplar. Therefore, in future revisions of the SLA, it would be important to develop alternate versions of these items or to find equivalent items that assess the same content.

Instructional Sensitivity. To investigate the instructional sensitivity of the SLA, we performed a series of ANOVAs and examined whether there were differences in the performance between School 1 (intervention group) and School 2 (comparison group)

on SLA-Total scores and the two SLA subscores. The results indicate that there was a statistically significant difference in performance between the two groups on the SLA-Total scores, $F(1, 98) = 44.10, p < .01$. The mean SLA-Total score for children in the intervention group (School 1) was 16. 91 ($SD = 3.74$). The mean SLA-Total score for the comparison group (School 2) was 12.03 ($SD = 3.07$). The effect size, as measured by Cohen's d ($d = 1.4$), was quite large (Cohen, 1988). These results are consistent with theoretical predictions that the intervention group would learn more science than the comparison group and indicate that the SLA is sensitive to variations in science instruction.

Our findings indicate that there were significant differences in performance between the two groups on both the SLA-Scientific Inquiry Processes, $F(1, 98) = 66.90, p < .01$, and the SLA-Life Science Concepts, $F(1,98) = 12.81, p < .01$, subscores. The effect sizes were large for both subscales (SLA-Scientific Inquiry Processes: Cohen's $d = 1.26$; SLA-Life Science Concepts*:* Cohen's $d = .84$), although larger on the items that assessed children's understanding of scientific inquiry processes. The mean SLA-Scientific Inquiry Processes subscore for children in the intervention group was 5.77 ($SD = 1.76$), indicating that on average they answered 5–6 of these 9 items correctly. In contrast, the mean subscores on these items for the children in the comparison group was 2.89 ($SD = 1.53$), indicating that on average these children answered fewer than 3 of the 9 items correctly. The differences on the SLA-Life Science Concepts subscores, although statistically significant, were smaller. The mean SLA-Life Science Concepts subscore for the intervention group was 11.14 ($SD = 2.75$) and the mean score for the comparison group was 9.14 ($SD = 2.28$).

Item Difficulty and Discrimination. An examination of item means and standard deviations (see Table 5) indicates that SLA items cover an adequate range of difficulty. The easiest SLA item was Item 23 (answered correctly by 95% of children); this item assessed their understanding of camouflage. The most difficult was Item 6 (answered correctly by 2% of children), which assessed

Table 5

SLA Item Difficulty and Discrimination: Mean, *SD*, and Discrimination Index

Scientific Inquiry Processes			
Item #	Mean	*SD*	Discrimination Index
1. Doing science	.52	.50	0.65
2. Science question	.27	.45	0.37
3. Observation—fish	.60	.49	0.69
4. Prediction—ball	.59	.49	0.83
5. Prediction—teeter totter	.47	.50	0.4
6. Hypothesis test	.02	.14	0.07
7. Science tools—record observations	.58	.50	0.65
8. Science tools—magnify	.81	.42	0.46
Life Science Concepts			
9. Science tools—measure temperature	.90	.30	0.14
10. Choose insect—easy	.92	.27	0.15
11. Justification for 10	.36	.48	0.79
12. Choose insect—difficult	.67	.47	0.61
13. Justification for 12	.58	.50	0.61
14. Caterpillar structure / function—eat	.73	.45	0.29
15. Caterpillar structure / function—breathe	.38	.49	0.69
16. Caterpillar structure / function—move	.83	.38	0.12
17. Butterfly life cycle	.88	.33	0.25
18. Camouflage—butterfly	.94	.24	0.14
Life Science Concepts			
Item #	Mean	*SD*	Item Discrimination
19. Not camouflaged—goldfish	.76	.43	0.29
20. Choose living thing—plant	.65	.48	0.54
21. Justification for 20	.45	.50	0.45
22. Living things need air	.85	.36	0.33
23. Living things need food	.95	.22	0.14
24. Identify invertebrate—crab	.49	.50	0.47

their understanding of hypothesis testing. On occasion, we also included easy and difficult items to test the same Life Science Concepts (e.g., Item 10–easy and 12–difficult) that address the characteristics of insects, and Items 18 (easy) and 19 (difficult) that address children's understanding of the concept of camouflage. Overall, the items assessing Scientific Inquiry Processes were harder than those assessing Life Science Concepts (see Table 5). To examine how SLA items discriminated between high and low scorers, we computed the Discrimination Index for each item (see Table 5) using Kelley's (1939) procedure. Following Ebel's (1954) criteria, 13 SLA items (1, 3, 4, 5, 7, 8, 11, 12, 13, 15, 20, 21, and 24) had moderate to high discrimination (ranging from .40 to .83). Eleven SLA items (2, 6, 9, 10, 14, 16, 17, 18, 19, 22, and 23) did not discriminate as well (Discrimination Index < .40) between high and low scorers compared to the former items. However, the average discrimination across all items was .43.

Correlations With Other Measures of Achievement and Science Learning. We obtained convergent validity information by correlating the children's scores on the SLA with external measures of science knowledge that included scores on WJ-III Science subtest and the SLP Portfolio Assessment. Discriminant validity information was obtained through correlations of the SLA with the WJ-III Passage Comprehension subtest.

SLA-Total scores and SLA-Life Science Concepts subscores were significantly associated with performance on the WJ-III Science subtest ($r = .38$, $p < .01$ and $r = .50$, $p < .01$, respectively). However, SLA-Scientific Inquiry Processes subscores did not correlate with the WJ-III science scores ($r = .11$). This is not surprising considering that the WJ-III Science subtest is a test of general knowledge and the items (for assessing young children) are not designed to measure children's understanding of the processes involved in scientific inquiry.

Further evidence for the validity of the SLA was obtained from correlations of the measure with SLP Portfolio scores, available for children in the intervention group. Composite Portfolio scores had a significant positive correlation ($r = .74$, p

< .01) with SLA-Total scores. Correlations also were computed for corresponding SLA and SLP Portfolio subscores. There was a significant positive correlation between the SLA-Scientific Inquiry Processes subscores and the Portfolio Inquiry Processes subscores (r = .57, p < .01). There also was a significant positive correlation between the SLA-Life Science Concepts subscores and the Portfolio Life Science Concepts scores (r = .58, p < .01). These results provide further convergent validity data to support the instructional sensitivity of the SLA.

There were relatively small correlations for the WJ-III Passage Comprehension subtest with the SLA-Total scores (r = .23 p = .02) and the SLA-Life Science Concepts subscores (r = .25, p = .02). The correlation was not statistically significant for the SLA-Scientific Inquiry Processes subscores (r = .14, p = ns). The fact that the SLA did not correlate as highly with the WJ-III Passage Comprehension subtest as it did with the WJ-III Science subtest provides discriminant validity evidence and supports the inference that performance on the SLA subscales does not depend heavily on children's verbal skills.

Discussion

The results described above indicate that the scores obtained on the SLA in the current sample manifest adequate psychometric properties with regard to score reliability and validity. The SLA is an instructionally sensitive measure of kindergarten children's science learning.

Our findings on item difficulty are consistent with developmental research that suggests that typically children do not develop an understanding of the nature of scientific inquiry in the absence of systematic instruction and that understanding aspects of scientific inquiry is particularly hard for younger children. For example, developmental research indicates that hypothesis testing is a particularly difficult concept for young children to learn (Klahr, 2000; Kuhn, Amsel, & O'Loughlin, 1988; Kuhn & Dean, 2005; Schauble, 1996). In contrast, our

results on SLA items that assess life science concepts indicate that kindergarten children are able to develop a rich knowledge base about living things.

Although the items that assess an understanding of scientific inquiry processes proved more difficult for the sample as a whole, the instructional sensitivity of the SLA allows us to document that SLP kindergarten students did know significantly more about scientific inquiry than a demographically similar group of kindergarten students who did not receive targeted science instruction. In other words, the results of the SLA indicate that even young children can begin to develop an understanding of scientific inquiry with appropriate instructional support. Our findings are consistent with recent work by researchers such as Metz (2004), who argue for a richer conception of children's developmental capacities in the context of science instruction.

An important consideration in the design and interpretation of on-demand assessments such as the SLA is the pragmatic context that motivates their development and use. Because such assessments often have been misused and misinterpreted, we want to raise certain considerations in the use and interpretation of the SLA. The SLA is not intended for use as a general test of science achievement. It is a research instrument that is specifically designed to evaluate concepts targeted in the SLP research intervention. What we hope is generalizable from the SLA is that this is an approach to assessing science learning that can be easily adapted to varied instructional contexts. For instance, our approach may be adapted to develop new items to test science concepts not covered in SLP instruction (e.g., earth science concepts).

Additionally, although we computed separate subscores to assess children's understanding of scientific inquiry and their understanding of life science concepts respectively, we do not conceive these two aspects of science knowledge as mutually independent components of science learning. We believe that while children may acquire knowledge about the natural world without necessarily understanding the processes by which such scientific knowledge is culturally constructed, the converse is unlikely.

An important consideration with on-demand assessments for young children is that they may underrepresent children's capacities for scientific reasoning and learning in more richly contextualized and socially supported environments. For example, Hammer and Elby (2002) noted that children's performance and reasoning varies with contextual epistemological resources such as task, teacher, and peer support. Our research on the SLP project (Samarapungavan et al., 2008) also indicates that children's "enacted epistemologies" (Chinn & Samarapungavan, 2005) or their functional understandings of scientific inquiry are more visible in the supportive context of inquiry oriented classroom instruction and discourse. Therefore, it is important to use a variety of assessments, including more immediate assessments of children's classroom learning such as portfolios, to fully understand young children's science learning.

Conclusions and Directions for Future Research

As educators develop new science curricula and programs to address the lack of rich and challenging science instruction in the early grades, there is a need to document what children learn from such efforts. In order to develop research-based and pedagogically effective science curricula, we need assessments with clearly described theoretical and psychometric characteristics. The SLA is one example of such an assessment that can be used to aggregate and compare learning outcomes, as well as to provide empirical information on kindergarten children's capacities for science learning. In future research, we plan to make further revisions to the SLA, based on the findings of the current study and to collect and analyze additional data on the psychometric properties of the SLA. Public schools in our state do not typically identify or provide special programs for academically gifted students at the kindergarten level, and the participating kindergarten classrooms in this study did not offer any programs for gifted and talented students. However, we have some initial

research that suggests that the SLA could be used in conjunction with the SLP Portfolio measure to identify high-achieving kindergarten science students (Tsai & Samarapungavan, 2007, 2009). Future research will further examine how the SLA might be used in conjunction with more proximal assessments such as the SLP Portfolio assessment to identify kindergarten science learners with special needs including very low- or very high-achieving science students. The general approach used in its development, administration, and interpretation can be extended to other instructional contexts in science.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.

American Association for the Advancement of Science. (1993). *Benchmarks for scientific literacy.* Retrieved February 11, 2007, from http://www.project2061.org/publications/bsl/online/ch1/ch1.htm

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13,* 186–203.

Boyd, R., & Richerson, P. J. (2005). *The origin and evolution of cultures.* Oxford, UK: Oxford University Press.

Brown, A. L. (1997). Transforming schools into communities of thinking and learning about serious matters. *American Psychologist, 32,* 399–413.

Carey, S., & Spelke, E. (1994). Domain specific knowledge and conceptual change. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind* (pp. 169–200). Cambridge, UK: Cambridge University Press.

Center for Science and Mathematics Education. (1996). *National science education standards.* Retrieved February 11, 2007, from http://books.nap.edu/html/nses/1.html

Chinn, C. A., & Samarapungavan, A. (2005, July). *Toward a broader conceptualization of epistemology in science education.* Paper presented at the meeting of the Eighth International History, Philosophy, Sociology, and Science Education Conference, University of Leeds, England.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Driver, R., Asoko, H., Leach, J., Mortimer, E., & Scott, P. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher, 23,* 5–12.

Ebel, R. T. (1954). Procedures for the analysis of classroom tests. *Educational and Psychological Measurement, 14,* 352–364.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269–314). Greenwood, CT: Information Age.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9,* 466–491.

French, L. (2004). Science as the center of a coherent, integrated early childhood curriculum. *Early Childhood Research Quarterly, 19,* 138–149.

Gelman, R., & Brenneman, K. (2004). Science learning pathways for young children. *Early Childhood Research Quarterly, 19,* 150–158.

Gitomer, G., & Duschl, R. A. (1998). Emerging issues and practices in science assessment. In B. J. Fraser & K. G. Tobin (Eds.), *International handbook of science education* (pp. 791–810). Boston: Kluwer Academic.

Greeno, J. G. (1998). The situativity of knowing, learning, and research. *American Psychologist, 53,* 5–26.

Hammer, D., & Elby, A. (2002). On the form of a personal epistemology. In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 169–190). Mahwah, NJ: Lawrence Erlbaum.

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

Indiana Department of Education. (2006). *Indiana's academic standards: Kindergarten science*. Retrieved February 11, 2007, from http://www.doe.state.in.us/standards/docs-Science/2006-Science-Grade0K.pdf

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30,* 17–24.

Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes.* Cambridge, MA: MIT Press.

Klein, E. R., Hammrich, P. L., Bloom, S., & Ragins, A. (2000). Language development and science inquiry: The Head Start on Science and Communication program. *Early Childhood Research and Practice, 2,* 1–22.

Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). *The development of scientific reasoning skills.* Orlando, FL: Academic Press.

Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science, 16,* 866–870.

Mantzicopoulos, P., Patrick, H., & Samarapungavan, A. (2005). The scientific literacy project. Grant proposal to the Institute of Education. Unpublished manuscript.

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

Meisels, S. J., Liauw, 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.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23,* 12–23.

Metz, K. E. (2004). Children's understanding of scientific inquiry: Their conceptualization of uncertainty in investigations of their own design. *Cognition and Instruction, 22,* 219–290.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide* (3rd ed.). Los Angeles: Muthén & Muthén.

National Association for the Education of Young Children. (2003). Executive summary. Early learning standards: Creating the conditions for success. *Young Children, 58,* 69–70.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.

Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context*. New York: Oxford University Press.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching, 39,* 369–393.

Samarapungavan, A., Mantzicopoulos, P., & Patrick, H. (2008). Learning science through inquiry in kindergarten. *Science Education, 92,* 868–908.

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology, 32*, 102–119.

Schrank, F. A., McGrew, K. S., & Woodcock, R. W. (2001). *Technical abstract* (*Woodcock-Johnson III Assessment Service Bulletin No. 2*). Itasca, IL: Riverside.

Sheppard, L. (2000). The role of assessment in a learning culture. *Educational Researcher, 29,* 4–14.

Shepard, L., Kagan, S. L., & Wurts, E. (1998) *Recommendations and principles for early childhood assessments*. Washington, DC: National Educational Goals Panel Report. (ERIC Document Reproduction Service No. ED416033)

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56,* 197–208.

Tsai, M., & Samarapungavan, A., (2007, October). *Understanding the characteristics of gifted kindergarten students in science learning*. Paper presented at the annual conference of the National Association for Gifted Children, Minneapolis, MN.

Tsai, M., & Samarapungavan, A. (2009, April). *Identifying the characteristics of gifted kindergarten students in science learning.* Paper presented at the annual conference of the American Educational Research Association, San Diego, CA.

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

Yu, C. Y., & Muthén, B. (2002). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes* (Tech. Rep.). Los Angeles: University of California at Los Angeles, Graduate School of Education & Information Studies.

Authors' Note

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305K050038 to P. Mantzicopoulos, H. Patrick, and A. Samarapungavan. The opinions expressed are those of the authors and do not represent views of the U.S. Department of

Education. Appreciation is expressed to Karleah Harris, Meng-Fang Tsai, Tyler Brown, Melissa Stewart, Lisa Duffin, Anna Strati, and Sybil Durand for their help in the data collection and preparation of materials for this study. We also greatly appreciate the involvement of the teachers and children in this study.