JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 48, NO. 9, PP. 1079–1107 (2011)

Measuring Knowledge Integration: Validation of Four-Year Assessments

Ou Lydia Liu,[1] Hee-Sun Lee,[2] and Marcia C. Linn[2]

[1]*Educational Testing Service, 660 Rosedale Rd, Princeton, New Jersey, 08541*
[2]*Graduate School of Education, UC Berkeley, California*

Received 22 November 2010; Accepted 28 August 2011

Abstract: Science education needs valid, authentic, and efficient assessments. Many typical science assessments primarily measure recall of isolated information. This paper reports on the validation of assessments that measure knowledge integration ability among middle school and high school students. The assessments were administered to 18,729 students in five states. Rasch analyses of the assessments demonstrated satisfactory item fit, item difficulty, test reliability, and person reliability. The study showed that, when appropriately designed, knowledge integration assessments can be balanced between validity and reliability, authenticity and generalizability, and instructional sensitivity and technical quality. Results also showed that, when paired with multiple-choice items and scored with an effective scoring rubric, constructed-response items can achieve high reliabilities. Analyses showed that English language learner status and computer use significantly impacted students' science knowledge integration abilities. Students who took the assessment online, which matched the format of content delivery, performed significantly better than students who took the paper-and-pencil version. Implications and future directions of research are noted, including refining curriculum materials to meet the needs of diverse students and expanding the range of topics measured by knowledge integration assessments.

Keywords: item format; item response theory; knowledge integration; science assessment; validation

There is a widespread interest in developing assessments that accurately capture student learning of complex science topics through inquiry-based instruction (Duschl, Schweingruber, & Shouse, 2007; Krajcik & Sutherland, 2010). The field needs more science assessments that are valid indicators of inquiry learning and also psychometrically sound. A particular challenge in assessment design is to capture inquiry-based learning while reliably differentiating student performances. In addition, assessments are valuable when they extend the curriculum by providing authentic opportunities for students to demonstrate what they know and can do.

This paper describes a 4-year study developing assessments to measure knowledge integration among middle and high school students. We started by characterizing a knowledge integration construct, which was used to align assessment development and the development of the scoring rubrics. The items went through many iterations of refinement. Each iteration consisted of content review, pilot testing, psychometric analysis, teacher interviews, and cohort comparisons. In the development of the items, we sought to meet several criteria including (a) taking advantage of the authenticity of constructed-response items while minimizing complexities in designing scoring rubrics for these items, (b) ensuring the instructional

Correspondence to: O.L. Liu; E-mail: lliu@ets.org
DOI 10.1002/tea.20441
Published online 20 September 2011 in Wiley Online Library (wileyonlinelibrary.com).

sensitivity of the assessments while attending to the generalizability of the findings, and (c) providing in-depth evaluation of student learning while meeting technical quality and efficiency standards for items.

Literature Review

To advance science education in the United States, both student characteristics and the assessments used to evaluate science learning play significant roles. The broad impact on science education brought about by the U.S.'s rapidly changing demographics and developing technologies needs to be recognized in understanding future directions of science education (Kirsch, Braun, Yamamoto, & Sum, 2007). Designing assessments to serve diverse student populations is of critical importance. For example, ELL students represent the fastest growing population in American public schools (Government Accountability Office, 2006). In 2004–2005, about 5.1 million students were English language learners in the U.S. (Payán & Nettles, 2006). Improving ELL students' science achievement is crucial to the success of science education in this country. Gender is also a frequently researched variable in science education. Females are seriously underrepresented in certain science-related fields such as physics (American Physical Society [APS], 2007). The unbalanced gender workforce in science has a deep root in K-12 science education. As science education should provide equal opportunities for all subgroups of students, we review the roles that gender, ELL status, and use of computers play in student science achievement. Use of computer was included for the examination of technological impact on science performance, and served as a proxy for socio-economic status (SES) to disentangle the confounding between ELL status and SES, as a large portion of ELL students come from socially and financially disadvantaged backgrounds. On the test side, we review the impact of assessment delivery mode (online vs. paper-and-pencil) and assessment format (multiple-choice items vs. constructed-response items) on student science learning.

Gender and Science Achievement

Although the science gender gap has been narrowing for the past 30 years, males still dominate certain science fields such as physics and engineering (Fadigan & Hammrich, 2004). Females only represented 21% of the undergraduate degrees in physics in 2005 (National Science Board, 2008). Females only constitute about 18% of the doctoral degrees granted in physics (APS, 2007). In 2006, among all the full professors in physics departments, only 6% were female (APS, 2007).

Gender difference on science performance is consistently documented on standardized tests. On the 2005 NAEP science assessment, males outperformed females across 4th, 8th, and 12th grades (National Center for Education Statistics [NCES], 2006). The percentage of male learners that achieve the level *Proficient* or above on the NAEP science scales in grades 4, 8, and 12 is also higher than the percentage of females that achieve the same level in these grades (NCES, 2006).

Although males and females showed no significant overall differences in performance on the Programme for International Student Assessment (PISA) science assessment, males significantly outperformed females in the content area ''Physical systems'' in all Organization for Economic Co-operation and Development (OECD) member countries except Turkey (OECD, 2007). The Physical systems content area assesses students' knowledge of structure of matter, properties of matter, motions and forces, and energy transformations (OECD, 2007). The finding is consistent with results from other research studies in which males have outperformed females in physical sciences (Brotman & Moore, 2008).

Various reasons have been proposed to explain the gender difference in science performance. Females tend to have more negative attitudes toward science and fewer science experiences than males (Brotman & Moore, 2008; Catsambis, 1995). Females also tend to choose fewer science courses (Farenga & Joyce, 1999) and have a lower self-concept in learning science. In the 2006 PISA, males rated their science ability significantly higher than females in 22 out of the 30 participating OECD countries (OECD, 2007). Although performing as well as males, high school females tend to report higher levels of science anxiety in learning physical science (Britner, 2008). In a study surveying more than 3,000 college students, females indicated significantly less experience with conceptual understanding in high school physics classes and less discussion of the benefits of being a physicist, both being significant predictors of one's perceived physics identity (Hazari, Sonnert, & Sadler, 2010).

English Language Learners and Science Learning

English language learners constitute the fastest growing population in U.S. schools (Kirsch et al., 2007). Many of the ELL students in the U.S. come from economically disadvantaged families (Adedi & Gandara, 2006). ELL students face a dual challenge in science learning, as they need to improve English proficiency and acquire science content knowledge at the same time. Research has shown that ELL students have consistently performed lower than non-ELL students on reading, mathematics, and science (Abedi, 2002). On the NAEP 2005 science assessment, ELL students scored significantly lower than non-ELL students across grades 4, 8, and 12 (NCES, 2006). At grade 8, only 3% of ELL students achieved the NAEP *Proficient* level as compared to 30% of non-ELL students (NCES, 2006). In a study comparing ELL and non-ELL performance in three states (Kim & Herman, 2009), results showed that the achievement gap in science could be as large as 1.1 standard deviations. Findings on the performance of former ELL students are mixed. De Jong (2003) reported that after exiting bilingual or English as a second language programs, former ELL students are still less likely to demonstrate proficiency in math and science than non-ELL students. However, the Kim and Herman (2009) study found that after controlling for free or reduced lunch, former ELLs performed better than or as well as non-ELLs. Data from New York City also showed that former ELLs outperformed non-ELLs on state English language and math tests (New York City Department of Education, 2009).

Teachers can play a key role in helping ELL students learn science (Lee, Maerten-Rivera, Penfield, LeRoy, & Secada, 2008). Teachers need support and training in understanding the factors responsible for ELL students' misconceptions and learning difficulties (Lee, Maerten-Rivera, et al., 2008). Teachers also need feedback from students and peers in understanding specialized needs of ELL students which is not part of a teacher's regular training (Buck, Mast, & Ehlers, 2005). Researchers have implemented large-scale professional development training programs to equip teachers with the knowledge and skills to meet the special needs of ELL students (Lee, Hart, Cuevas, & Enders, 2004; Lee et al., 2008; Lee, Penfield, & Maerten-Rivera, 2009). Researchers have also found that inquiry-based science learning activities provide authentic experiences for ELL students and contribute to the effective communication of science understanding (Lee, Deaktor, Hart, Cuevas, & Enders, 2005).

Computer Use and Science Achievement

Studies have documented a positive relationship between computer use and student science learning (Christmann & Badgett, 1999; James & Lamb, 2000; Taningco & Pachon, 2008). For example, the achievement gap between Hispanic and White fifth graders on science could be as large as 0.35 standard deviations (NCES, 2004). However, after

controlling for variables including computer use, the performance gap is narrowed to 0.10. Students who use computers during science class almost every day significantly outperform students who rarely use computers in science class (Taningco & Pachon, 2008).

There are also mixed findings on computer use and academic achievement. Wenglinsky (1998) reported that students who spent more time on computers in school actually performed less well than students who spent less time. The effect of computer use on academic achievement could depend on how computers are used. The Trends in International Mathematics and Science Study (TIMSS) has also documented a negative correlation between computer use and student learning. Students from Cyprus, Hong Kong, and the United States have been shown to have the lowest achievement when they use computers frequently in class (Papanastasiou, 2002). To further clarify the impact of computer use on student science literacy, Papanastasiou, Zembylas and Vrasidas (2003) studied the U.S. participants in the 2000 PISA science assessment. After student socioeconomic status was controlled for, computer availability at home and at the library was significantly correlated with higher science achievement. Student comfort levels with using a computer to write papers and student use of computers for educational purposes have also been associated with success in science learning (Papanastasiou et al., 2003). In sum, computer use could have a positive impact on student achievement, when computers are used appropriately for learning purposes.

Online Versus Paper-and-Pencil Assessment

With the rapid development of educational technology, online testing has replaced traditional paper-and-pencil testing in many settings. As the *Race to the Top* (RTTT; 2009) initiative calls for large-scale anchor assessments for high school students that are aligned with college and career readiness standards, online testing becomes a viable choice for effective assessments. Online testing has the advantage of easy administration, flexible scheduling, easy data access, and efficiency (Bodmann & Robinson, 2004; Gallagher, Bridgeman, & Cahalan, 2002). Although the literature on comparing the two testing delivery formats is relatively sparse in science education, abundant literature exists documenting such a comparison in other fields, such as math, language tests, army intelligence tests, and undergraduate or graduate admissions tests (Educational Testing Service, 2005; Gallagher et al., 2002; Segall, 1997). Most of the previous investigations focus on understanding whether there are group differences when switching from paper-and-pencil tests to online tests, and if so, what are the scope and sources of such differences. A general finding is that the magnitude of performance difference varies across subject domains (e.g., math, verbal) by gender and ethnicity groups between the paper-and-pencil and online tests, and the differences tend to be small (Bodmann & Robinson, 2004; Gallagher et al., 2002). Kingston (2009) conducted a meta-analysis with 81 studies on the comparability between computer-based and paper-and-pencil multiple-choice tests in K-12 and found that the average effect size was very small across subjects.

Besides studying the impact of student characteristics and delivery format on science performance, we also need to consider what forms of assessment are most valid and effective in measuring science achievement for students from all backgrounds. Current science education standards (National Research Council, 1996) emphasize the importance of using a chain of reasoning that connects student performance data to assertions made about student learning through inquiry. Few assessments provide this chain of reasoning explicitly. Challenges include the design of items and the design of the scoring rubric. To measure science learning emerging from inquiry-based instruction, items should provide opportunities for students to formulate ideas, support or refute an explanation of scientific phenomena, and provide justifications for their explanations. In addition to these diagnostic and learning requirements, items

and scoring rubrics need to meet technical and psychometric requirements. Since science achievement has often been measured with multiple-choice and constructed-response items, we review the roles that each assessment format plays in assessing inquiry-based science learning.

Multiple-Choice Items

Using multiple-choice items can be an efficient method for assessing minimum competency and basic scientific knowledge and skills. Basic definitions and scientific facts can be considered basic knowledge for students to perform more complex scientific investigations later. Multiple-choice items are well suited to measure discrete pieces of knowledge given their efficiency and relative ease of scoring (Clark & Linn, 2003; Liu, Lee, & Linn, 2010a, 2010b). Because of their high reliability, objectivity, and straightforward scoring rubrics, multiple-choice items are widely used in large-scale standardized assessment. For example, in the released 2003 TIMSS 4th grade science assessment, 50 of the test's 73 items were multiple-choice (TIMSS, 2003).

Despite these advantages, multiple-choice items often fall short in assessing deep scientific reasoning and explanation (Smith, Wiser, Anderson, & Krajcik, 2006; Stern & Ahlgren, 2002). Multiple-choice items are not well suited to capturing the complex nature of science learning through inquiry. They also have difficulty capturing the different approaches students take in solving scientific problems (Alonzo & Steedle, 2009). Researchers have explored alternative multiple-choice item formats in an attempt to enhance their functionality in measuring complex science learning. For example, Briggs, Alonzo, Schwab, and Wilson (2006) designed ordered multiple-choice items where the choices represent different levels of scientific understanding. Sadler (1998) used the multiple choices as distracters to reveal student misconceptions related to a given scientific phenomenon. These research efforts are important first steps towards assessing the reasoning underlying student responses to multiple-choice items.

However, these multiple-choice items do not allow for direct inferences regarding student reasoning, as student ability to select among choices is different from the ability to generate justifications on their own. Liu, Lee, and Linn (in press) designed a type of multiple-choice item called explanation multiple-choice (EMC), in which the choices represent different levels of scientific reasoning and are created based on student responses to previously administered constructed-response items. The EMC questions have six choices and each choice corresponds to a level on the knowledge integration construct. Liu, Lee, and Linn compared student performances on EMC items and regular constructed-response items where students generate their own explanations. The results showed that it is more difficult for students to explain science phenomena using their own terms than to choose from a set of provided explanations.

Constructed-Response Items

Constructed-response items provide students with opportunities to express their own ideas and solve science problems in their own terms (McCarthy, 2005). Compared to answering multiple-choice items, students tend to panic less when providing written explanations to constructed-response items (Rockow, 2008). Constructed-response items also provide students the chance to choose a position and fully elaborate on their position using supporting evidence, an activity which is similar to scientific reasoning and critical thinking in real-life science inquiries.

Although constructed-response items have great potential in eliciting students' ideas and approaches, they take longer for students to complete than multiple-choice items (Thissen, Wainer, & Wang, 1994). They also require a clearly defined and fully elaborated scoring rubric to identify students' ideas and thinking. The development of the scoring rubric is as important as the development of the items, because the rubric determines the kind of criteria used to evaluate student responses and, therefore, also determines whether the scores reflect the intentions of the item design.

Large-scale standardized assessments such as the TIMSS and PISA often include constructed-response items. However, their scoring rubrics often fail to distinguish among the levels of knowledge and reasoning targeted by the items. For example, there are several constructed-response items in the 2003 TIMSS 4th grade science assessment (TIMSS, 2003). These items either ask the students to explain their answers to a previous multiple-choice item or to describe the differences between two scientific processes or substances. However, most of the time, the scoring rubrics are designed to capture only correct and incorrect answers, failing to consider the levels of understanding that exist between right and wrong answers.

Another potential caveat with respect to constructed-response items concerns the language demands they place upon the test taker. Due to the changing demographics of American K-12 education, an increasing number of students are English language learners (Kirsch et al., 2007). Special attention should be given to language learners in the use of constructed-response items so that the students are not unfairly disadvantaged by the language requirements of the items (Solano-Flores & Nelson-Barber, 2001).

Because of the subjectivity involved in scoring constructed-response answers, it is also important to ensure that the raters understand the scoring rubric sufficiently and are able to assign scores consistently across items and students. Inter-rater reliability is a primary concern for constructed-response items (Johnson, McDaniel, & Willeke, 2000). Wang and Wilson (1996) reported that the difference in student ability estimate due to the subjectivity of raters could be as large as 0.56 out of a 3.5 logit scale. Because of the complexities involved in scoring open responses, constructed-response items are more costly than multiple-choice items to administer and score.

The Purposes of the Study

The purposes of this study are to develop a set of valid and reliable assessments to measure knowledge integration and to study the relationship between various student and assessment characteristics and science performance. Three research questions are addressed in this study:

1. What is the technical quality of the knowledge integration assessments administered over 4 years?
2. How do student gender, ELL status, and use of computers affect students' science achievement?
3. How does test delivery format (online vs. paper-and-pencil) affect student science achievement?

This study was conducted as part of the research activities of Technology Enhanced Learning of Science, a large science teaching and learning research center funded by the National Science Foundation. A knowledge integration construct was established based on a learning theory called *knowledge integration* and was consistently used to guide the design of

the items and scoring rubrics and the interpretation of assessment results. With the intention of balancing items in terms of validity, reliability, format, and proximity to instruction, the knowledge integration assessments were composed of items from both published standardized tests and project-produced tests. In order to measure student reasoning, most of the multiple-choice items in the knowledge integration assessments are followed by a constructed-response ''Explain your choice'' item. In the following section, we describe the knowledge integration construct, the design of the items, test implementation, psychometric properties of the items, and how the knowledge integration assessments serve students of different characteristics with regard to gender, ELL status, and computer use.

The Knowledge Integration Construct

Knowledge integration theory provides design rationale for both curriculum modules and assessment (Lee, Linn, Varma, & Liu, 2010). Knowledge integration is a theory that represents cognition in terms of multiple, diverse, and often contradictory ideas students have about scientific phenomena and links students make among these ideas (Linn, 1995, 2006; Linn, Davis, & Bell, 2004; Linn & Hsi, 2000; Linn, Lee, Tinker, Husic, & Chiu, 2006). From the knowledge integration perspective, science learning occurs when students elicit their own ideas, add new normative ideas from formal and informal science instruction, develop scientific criteria to distinguish between the ideas, and form more coherent views of science as a result of integrating various scientific ideas. For the design of curriculum modules, research identified four knowledge integration design principles such as making thinking visible, making science accessible, learning from each other, and supporting lifelong learning (Linn & Eylon, 2006). The curriculum modules were implemented through a web-based learning environment called the Web-based Inquiry Science Environment (WISE, http://wise.berkeley.edu). WISE incorporates technological features and provides opportunities for students to experience interactive visualizations, in-class experiments, collaborative learning, embedded assessments and real-time feedback. The WISE features allow students to have direct observations of science phenomena and develop understanding of the connections between different phenomena through manipulation of variables.

For the design of assessment, knowledge integration suggests directions for how science learning might occur to students:

- from not eliciting relevant ideas to eliciting science-related ideas
- from eliciting non-normative science ideas to normative ideas
- from eliciting normative science ideas to making a scientifically valid link among the normative science ideas elicited
- from making one scientifically valid link to making multiple scientifically valid links.

This set of the knowledge integration directions is used to characterize distinctive cognitive levels in the knowledge integration construct (Figure 1). While typical science assessment such as NAEP and TIMSS measure the understanding or knowledge of a particular science content domain, the knowledge integration assessments focus on a psychological attribute related to students' ability to elicit and connect scientific ideas.

In modeling the knowledge integration construct, we adopted the construct modeling approach put forward by Wilson (2005) where (1) the construct of interest can be represented on a single continuum, (2) students with more and less of the latent ability specified by the construct can be mapped onto the continuum, (3) items that measure more and less of the construct ability can be mapped onto the same continuum, and (4) a learning theory provides

Knowledge integration level (score)	Criteria	Examples
No answer (0)	Blank or random answer	• I don't know
Off-task (1)	Irrelevant to the question	
No-Link: Non-normative ideas (2)	Elicited non-normative ideas or restated the multiple choice answer.	• The elements go back to the environment *[restatement of the question and answer…no additional info is added]*.
Partial-Link: Normative ideas (3)	Elicited one of the ideas listed above.	• I believe that the animals and plants go inside of the ground once they die and decomposers feed on them *[process idea]*. • Because the elements go back in the earth. *[outcome idea]* • When things die they decompose. *[outcome idea]* • When they die they fertilize the soil. *[impact idea]*
Full-link: Single link between two normative ideas (4)	Used one of the links listed above.	• They get decomposed by a decomposer *[process + outcome]*. • The decomposers would put the nutrients back into the ground *[process + impact]*. • When an animal or a plant dies, the worms eat it and make it into soil, which trees need to survive *[process + impact]*.
Complex-Link: Two or more links between normative ideas (5)	Used the Process-Outcome-Impact link	• The dead animals and plants are decomposed (or change into elements) by decomposers and the elements are used by plants to grow *[process + outcome + impact]*

Figure 1. Scoring rubric for the Element constructed-response item.

a rationale for the placements of students and items on the continuum. In our case, the learning theory is the knowledge integration theory. The knowledge integration construct consists of six-distinct levels as shown in the first column of Figure 1.

The knowledge integration construct is intended to provide the continuum to which a student's ability to elicit and connect scientifically normative ideas can be mapped. The knowledge integration construct is intended to cover a full range of scientific understanding, also identified by other researchers, such as non-commitment to science, non-normative ideas (diSessa, 1993; Linn, 2006), normative but isolated ideas, to an integrated network of normative ideas similar to the structure of expert knowledge (Baxter & Glaser, 1998; Bransford, Brown, & Cocking, 2000; Chi, Feltovich, & Glaser, 1981).

To characterize student responses to these knowledge integration levels, we examined students' open-ended responses to explanation items. The National Science Education Standards (NRC, 1996) recognize that ''because explanation is central to the scientific enterprise, eliciting and analyzing explanations are useful ways of assessing science achievement''

(p. 92). The importance of explanations in science learning is well documented (Chi, 1998; McNeill, Lizotte, Krajcik, & Marx, 2006; Sandoval, 2003). For any given student's open-ended response to an explanation item, we examined the number of normative and relevant ideas elicited and the number of scientifically valid links among the elicited normative and relevant ideas (Lee, Liu, & Linn, 2011). The following is a sample item pair called ''Element'' with a multiple-choice item followed by a constructed-response item.

Sample Item—Element

Animals and plants are made up of a number of different chemical elements. What happens to all of these elements when animals and plants die?

a.They die with the animal or plant.
b.They evaporate into the atmosphere.
c.They are recycled back into the environment.
d.They change into different elements.

Explain your choice.

Here are the ideas that students can use to explain their response to the multiple-choice item:

- Process idea: Decomposition by decomposers (e.g., worms, bacteria)
- Outcome ideas: When plants/animals die, they are decomposed, or elements are released into the earth
- Impact ideas: Plants use the elements released, or some elements are used as fertilizer

Here are the links between ideas that students can draw on to justify their explanations:

- Process-Outcome link
- Process-Impact link
- Outcome-Impact link

A scoring rubric for the constructed-response part of the item pair is provided in Figure 1. The scoring rubric shows six knowledge integration levels with descriptions and student response examples. The knowledge integration rubric examines students' elicitation and use of scientific ideas in explaining a scientific phenomenon or in justifying a claim made about a scientific problem. Higher scores on the knowledge integration rubric represent more coherent views of a scientific phenomenon based on relevant scientific ideas rather than fragmented ideas.

In prior research, we developed knowledge integration scoring rubrics for 122 constructed-response items addressing physical, life, and earth science topics (Lee et al., 2010). While the content in each item is different, the general scoring approach is identical for all constructed-response items. Simplifying the scoring procedure helps achieve high inter-rater reliability (Liu, Lee, Hofstetter, & Linn, 2008). We applied an item response theory analysis based on the Rasch partial credit model (PCM; Rasch, 1960/1980; Wright & Masters, 1982) to the responses of 30,000 students to 180 multiple-choice items and 122 open-ended

explanation items. The results showed that the levels on the knowledge integration construct are distinctive, valid, and reliable (Liu et al., 2008) and that student responses to the multiple-choice items can be also interpreted on the same knowledge integration construct (Lee et al., 2011).

Participants

Students (18,729) were tested using the knowledge integration assessment at the end of each school year from 2005 to 2008 (Table 1). The students were taught by 340 teachers from 50 schools and 21 school districts in five states. In general, middle school students were tested in physical, life, or earth sciences, while high school students were tested in physics, chemistry, or biology. The total number of unique teachers was 228, as some teachers taught more than one area of science or taught the same science unit in more than 1 year. The students consisted of 56.6% middle school and 43.4% high school students; 49.7% were female; 21.5% spoke English as a second language; and 59.1% reported using computers for homework. Among the participants, about 43% took the assessment online, and the rest took the paper-and-pencil version. Assignment to the two-test delivery formats was random within the group of students taught by a given teacher.

Curricula and Assessment

From 2003 to 2008, we developed ten 1-week-long science units covering six science subjects including middle school life science, physical science and earth science, and high school physics, biology, and chemistry. These units target traditionally hard-to-teach topics such as global warming, airbags, and chemical reactions. Each unit takes advantage of simulation and visualization tools to allow students to directly observe the scientific phenomena. Students also have opportunities to interact with computer simulations by manipulating variables and observing the outcomes. As described earlier, all the learning activities took place in the WISE system. Since the knowledge integration assessments were administered to students at the end of the school year, there were delays ranging from a few weeks to months between the enactment of an inquiry-based unit and the administration of assessments.

Proximity to Instruction

In assessment development, we considered both the alignment between assessment and instruction and the generalizability of the results. Both proximal and distal items were

Table 1
Number of participating students and teachers by year and subject

	2005		2006		2007		2008		Total	
	n(S)	*n*(T)	*n*(S)	*n*(T)	*n*(S)	*n*(T)	*n*(S)	*n*(T)	*n*(S)	*n*(T)
MES	1236	14	1000	15	1113	15	780	11	4129	55
MLS	978	11	1396	16	993	13	861	12	4228	52
MPS	654	7	1129	16	1197	14	716	9	3696	46
PHY	372	9	128	4	63	4	39	1	602	18
CHM	529	11	627	10	548	10	312	5	2016	36
BIO	744	14	1445	56	1465	35	404	28	4058	133
Total	4513	66	5725	117	5379	91	3112	66	18729	340

MES, middle school earth science; MLS, middle school life science; MPS, middle school physical science; PHY, physics; CHM, chemistry; BIO, biology. *n*(S), number of students and *n*(T), number of teachers.

included. The proximal items were closely aligned with what students learned in the inquiry-based unit materials and were used to evaluate the impact of the target instruction. The design of the proximal items was guided by the knowledge integration theory, in that the items capture the levels of knowledge integration ability specified by the construct. The distal items were selected from published standardized science tests such as the TIMSS and NAEP. These items cover similar science content but are not directly related to the topics students learned in the inquiry science units (Liu et al., 2010a, 2010b).

Item Format

Three main item formats appear in the knowledge integration assessment: multiple-choice items, short-response items, and constructed-response items. In half of the cases, a multiple-choice item is followed by a constructed-response item as a two-tier item bundle. For example, after students respond to a multiple-choice item, they are asked to "explain your choice.." Our wide use of the two-tier item format was to keep a balance between efficiency and in-depth investigation of student reasoning. By combining these two, the knowledge integration assessments use multiple-choice items to scaffold student answers to the constructed-response component, and then use the student-generated responses to directly infer student reasoning. Prior research shows that the two-tier items are a practical and reliable item format for eliciting student explanations within a limited testing time (Lee et al., 2011). See Table 2 for the item composition by format and year.

Assessment Refinement

From 2005 to 2008, the knowledge integration assessments were given to four cohorts of students. While the major bulk of the items remained the same, the assessments went through a number of iterations for refinement. During the item development, draft items were reviewed by content experts, science educators, assessment specialists, researchers, and teachers. Items that were considered unclear, too easy or too difficult were either improved or removed from the item bank. Student think-alouds were also conducted for some of the constructed-response items to ensure the clarity of the questions and the appropriateness of the wording. After the items were pilot tested, item response models were used to evaluate the psychometric properties of the items in terms of item difficulty, item fit, person fit, test reliability, person reliability, and differential item functioning with regard to gender (Liu et al., 2008). Items identified with problems were carefully reviewed, revised or removed.

Table 2
Item distribution by format and year

	2005			2006			2007			2008			
	MC	SP	CR	MC	SP	CR	MC	SP	CR	MC	SP	CR	Total
MES	5	—	9	6	—	8	8	—	7	6	—	5	54
MLS	7	1	9	7	1	8	10	—	9	7	—	7	66
MPS	11	5	8	10	2	7	14	2	10	7	2	5	83
PHY	12	—	11	13	3	9	12	—	8	11	—	7	86
CHM	8	—	5	14	—	8	12	—	7	11	—	7	72
BIO	10	—	8	8	1	8	7	—	8	7	—	8	65

MC, multiple-choice items; SP, short response items; CR, constructed-response items; MES, middle school earth science; MLS, middle school life science; MPS, middle school physical science; PHY, physics; CHM, chemistry; BIO, biology.

The purposes of collecting 4 years' worth of data were to validate the items with different samples of students, to ensure a reasonably large sample size, to improve assessment quality, and to examine the stability of findings.

Assessment Continuity

In this study, knowledge integration assessments were used for students in different cohorts across years. Due to limited testing time, multiple test forms were used. Common items were embedded across multiple test forms, and test-equating techniques were used to ensure the comparability of the performance of students taking different test forms (Liu et al., 2010a, 2010b). Similarly, to address yearly differences in the item composition across assessments, a set of common items was embedded in tests across years. Student performances across years were equated on the basis of the common items using the mean/sigma equating method (Kolen & Brennan, 2004). Test equating techniques can be very useful in cases where different test forms are involved in performance comparison.

Scoring Rubrics

Knowledge integration scoring rubrics consistently reward student ability to generate relevant scientific ideas and connect the scientific ideas in explaining a scientific phenomenon or justifying a scientific claim. Since each item features a different scientific phenomenon or problem, the ideas and links needed to answer each item vary. Therefore, the knowledge integration scoring rubrics were developed to address specific features of each item. Each knowledge integration scoring rubric characterizes student responses into the six knowledge integration ability levels (Figure 1). Scores generated from these knowledge integration scoring rubrics are critical to the validity of the test scores. Since the constructed-response items measure important aspects of inquiry-based learning, such as scientific explanations and argumentations, it is particularly important for the scoring rubrics to differentiate among the kinds of student reasoning required for explaining the scientific phenomena and the types of evidence students use to justify their argumentations.

The knowledge integration scoring rubrics were intentionally designed to align with the levels specified in the knowledge integration construct. This alignment allows for comparisons of scores on the knowledge integration construct across different constructed-response items. For example, a score of 4 on two different items has the same meaning: the student holds the same level of knowledge integration ability with the topics addressed in the two items.

Analysis

In this study, both Rasch models and classical testing methods were used to examine the psychometric properties of the multiple-choice, short-answer, and constructed-response items administered across the 4 years. Specifically, we examined test reliability, distributions of item difficulty and student ability estimates, item fit, and person reliability.

The Rasch model (Rasch, 1960/1980), was used to analyze the multiple-choice and short-answer items. In the Rasch model, the probability of a correct answer on a multiple-choice item depends on both the item difficulty and the ability of the student. The easier the item and the more able the student, the more likely the student is to score 1 instead of 0 on the multiple-choice item, and vice versa. The Rasch model provides an estimate of difficulty for each dichotomously scored item and an estimate of ability for each student. The item and the student estimates are scaled to the same logit scale, usually ranging from -3 to 3. The larger the values, the more difficult the item is and the more able the student is. Rather than

treating the raw score as the true student score, the Rasch model provides an estimate of error for each of the ability estimates, called the standardized error of measurement (SEM). SEM provides information on how accurate the estimate is depending on the consistency of the student responses across items.

The Rasch PCM(Masters, 1982) was used to analyze the constructed-response items. The constructed-response items are scored as 0–5 using the knowledge integration scoring rubrics. In addition to providing an estimate of overall item difficulty, the PCM provides an estimate of difficulty for each of the steps, indicating the difficulty of scoring $k + 1$ instead of k (in this case, k takes values from 0 to 4) on a given constructed-response item. The PCM was selected rather than a rating scale model was because some categories on certain items may have very few observations. For example, not many students were able to achieve a score 5 on difficult constructed-response items. Another reason for choosing the PCM over the rating scale model is that the difficulty of going from score k to $k + 1$ may vary across items, which violates the assumption of the rating scale model that the scale structure of all items should be the same.

As mentioned earlier, to compare student performance across test forms and across years, test-equating techniques were applied based on a set of common items. For example, in 2005, two forms were designed for the middle school life science test, with 12 items in form A and 11 items in form B. These two forms share 10 common items. The estimated item difficulties of the common items on form A were transformed to have the same mean and standard deviation as the difficulties of the common items in form B. Based on the linear function, the student ability estimates were transformed to be on the same scale to allow comparison between forms (Kolen & Brennan, 2004).

Besides examining the psychometric properties of the assessment items across 4 years, analyses were conducted to investigate the impact of various student characteristics on science knowledge integration ability. Student variables including gender, language, and use of computers for homework were examined. Number of science and math courses taken was also included for high school students. In addition, we explored the relationship between teachers' implementation of the inquiry units and the knowledge integration performance of their students.

Results

Descriptive Statistics

ConQuest provides an ability estimate for each student with a distinct response vector. The estimates were placed on a logit scale, often ranging from -3 to 3. The larger the estimate, the more able the student is. Table 3 presents the mean student ability estimate by subject and by year. As each year a different cohort of students took the assessment, students from 1 year are not expected to score higher or lower than students from another year.

Test Reliability

The reliability indicated by Cronbach's alpha (Table 4) is satisfactory for all of the tests in the six subjects from 2005 to 2008. The Cronbach's alphas range from 0.72 to 0.84, with mean 0.79 and standard deviation 0.04.

In addition to the overall test reliability, analyses were also run to calculate the reliability of the constructed-response items (Table 5). Contrary to the commonly held notion that constructed-response items tend to have lower reliability than multiple-choice items, the constructed-response items in this study actually had fairly high reliabilities, with Cronbach's alpha ranging from 0.69 to 0.90, with mean 0.81 and standard deviation 0.05.

Table 3
Mean, standard deviation, and sample size by year and module

	2005			2006			2007			2008		
	M	SD	*n*	M	SD	*n*	M	SD	*n*	M	SD	*n*
MES	0.07	0.56	1236	0.05	0.57	1000	−.24	0.51	1113	0.16	0.60	780
MLS	0.03	0.78	978	−0.07	0.76	1396	0.02	0.66	993	0.06	0.70	861
MPS	0.05	0.62	654	−0.07	0.67	1129	−0.03	0.67	1197	0.10	0.58	716
PHY	0.04	0.70	372	0.14	0.71	128	−0.47	0.78	63	−0.17	0.54	39
CHM	0.06	0.88	529	0.09	0.73	627	−0.29	0.77	548	0.22	0.62	312
BIO	0.16	0.73	744	−0.07	0.79	1445	−0.05	0.72	1465	0.15	0.73	404

MES, middle school earth science; MLS, middle school life science; MPS, middle school physical science; PHY, physics; CHM, chemistry; BIO, biology.

Distribution of Item Difficulty and Student Ability Estimates

As described above, in Rasch-type models, the probability of a correct answer on an item depends on both the item difficulty and student ability. The ConQuest software (Wu, Adams, Wilson, & Haldane, 2007) provides an estimate of item difficulty for each item and an ability estimate for each student. Both item and student estimates are scaled to the same logit scale (often from −3 to 3) and, therefore, are comparable in value. The larger the value, the more difficult the item, and the more able the student. ConQuest also produces a graph referred to as the Wright map. The Wright map illustrates the distribution of item difficulty and student ability estimates and is used to evaluate both the internal design and scoring of the items and the overall performance of the students. Take middle school life science as an example (Figure 2). The numbers in the far left column represent the logit scale from −3 to 3. The next column contains the student distribution, with each "X" representing about 30 students. The position of the bars in the distribution suggests the value of the ability estimate. The higher the position, the more able the students are. For example, at the highest logit level of 3, the bar contains only one X (i.e., about 30 students), indicating the individuals who were the highest performing students on the middle school life science test.

Table 4
Number of items and reliability of the items by year and module

	2005		2006		2007		2008	
	No. of items	α	No. of items	α	No. of items	α	No. of items	α
MES	13[a]	0.72[b]	14	0.82	17[a]	0.77[b]	13	0.81
MLS	13[a]	0.78[b]	16	0.84	17[a]	0.76[b]	14	0.80
MPS	19[a]	0.72[b]	21	0.81	21[a]	0.76[b]	17	0.73
PHY	18[a]	0.81[b]	25	0.82	20	0.84	18	0.79
CHM	13[a]	0.80[b]	22	0.82	16	0.86	19	0.79
BIO	13[a]	0.72[b]	16	0.82	15	0.82	15	0.82

α is Cronbach's alpha.

MES, middle school earth science; MLS, middle school life science; MPS, middle school physical science; PHY, physics; CHM, chemistry; BIO, biology.

[a]since two test forms were used for each of the six subject areas in 2005 and for each of three middle school subject areas in 2007, the number of items in Table 4 is the number of unique items between the two forms.

[b]the reliability is the mean reliability between the two tests forms.

Table 5
Number of items and reliability of the CR items by year and module

	2005		2006		2007		2008	
	No. of items	α	No. of items	α	No. of items	α	No. of items	α
MES	9[a]	0.74[b]	8	0.83	10[a]	0.77 [b]	7	0.79
MLS	9[a]	0.85[b]	9	0.87	9[a]	0.79[b]	8	0.82
MPS	8[a]	0.69[b]	7	0.83	8[a]	0.80[b]	6	0.71
PHY	12[a]	0.81[b]	9	0.85	8	0.87	7	0.78
CHM	7[a]	0.83[b]	8	0.87	8	0.90	7	0.83
BIO	9[a]	0.76[b]	7	0.84	8	0.81	8	0.82

α is Cronbach's alpha.

MES, middle school earth science; MLS, middle school life science; MPS, middle school physical science; PHY, physics; CHM, chemistry; BIO, biology.

[a]Since two test forms were used for each of the six subject areas in 2005 and for each of three middle school subject areas in 2007, the number of items in Table 5 is the number of unique items between the two forms.

[b]the reliability is the mean reliability between the two-test forms.

The columns to the right of the student ability distribution represent the items. The numbers under the MC column are the multiple-choice items, and the numbers under the headings ''Threshold 1'' through ''Threshold 5'' are the item thresholds for the constructed-response items. For multiple-choice items, an estimate is provided to indicate the difficulty of that item. On the logit scale, the higher the position, the more difficult the item is. For example, among all multiple-choice items, item 2 is the most difficult one, and item 4 is the easiest.

As mentioned earlier, the constructed-response items are scored using a 0–5 scoring rubric. The thresholds are used to indicate the difficulty level of scoring $k + 1$ instead of k ($k = 0, 1,\ldots 4$). For example, 31.1 is threshold 1 for item 31 and 6.4 is threshold 4 for item 6. The higher the threshold, the more difficult it is to obtain the score corresponding to that threshold. The Wright map shows an expected progression of threshold difficulties from thresholds 1–5, with threshold 5 at the highest positions, followed by thresholds 4, 3, 2, and 1. The progression shows that the empirical data confirmed the design and scoring of the items. Scores of 5 were the most difficult to obtain, followed by scores of 4 and below. A comparison between the multiple-choice items and the constructed-response items showed that it is much more difficult to obtain a maximum score on the constructed-response items than obtaining the highest score on the multiple-choice items, as the logit values of the thresholds 5 on the constructed-response items are well above the logit value of even the most difficult multiple-choice item (i.e., item 2). A more detailed analysis on how multiple-choice and constructed-response items across all six subject areas contributed to the measurement of knowledge integration can be found in Lee et al. (2011).

Item Fit

ConQuest provides two kinds of fit statistics for an item: unweighted mean square fit (UMSF) and weighted mean square fit (WMSF). The UMSF is also known as the outfit statistic and the WMSF is also known as the infit statistic. The outfit statistic is sensitive to unexpected student responses that are far below or above their ability estimates. The infit statistic is sensitive to unexpected patterns of student responses across items. As the outfit statistic is easier to detect and manage than the infit statistic, it is used to indicate the fit between the observed item responses and the model expectation in this study. An acceptable

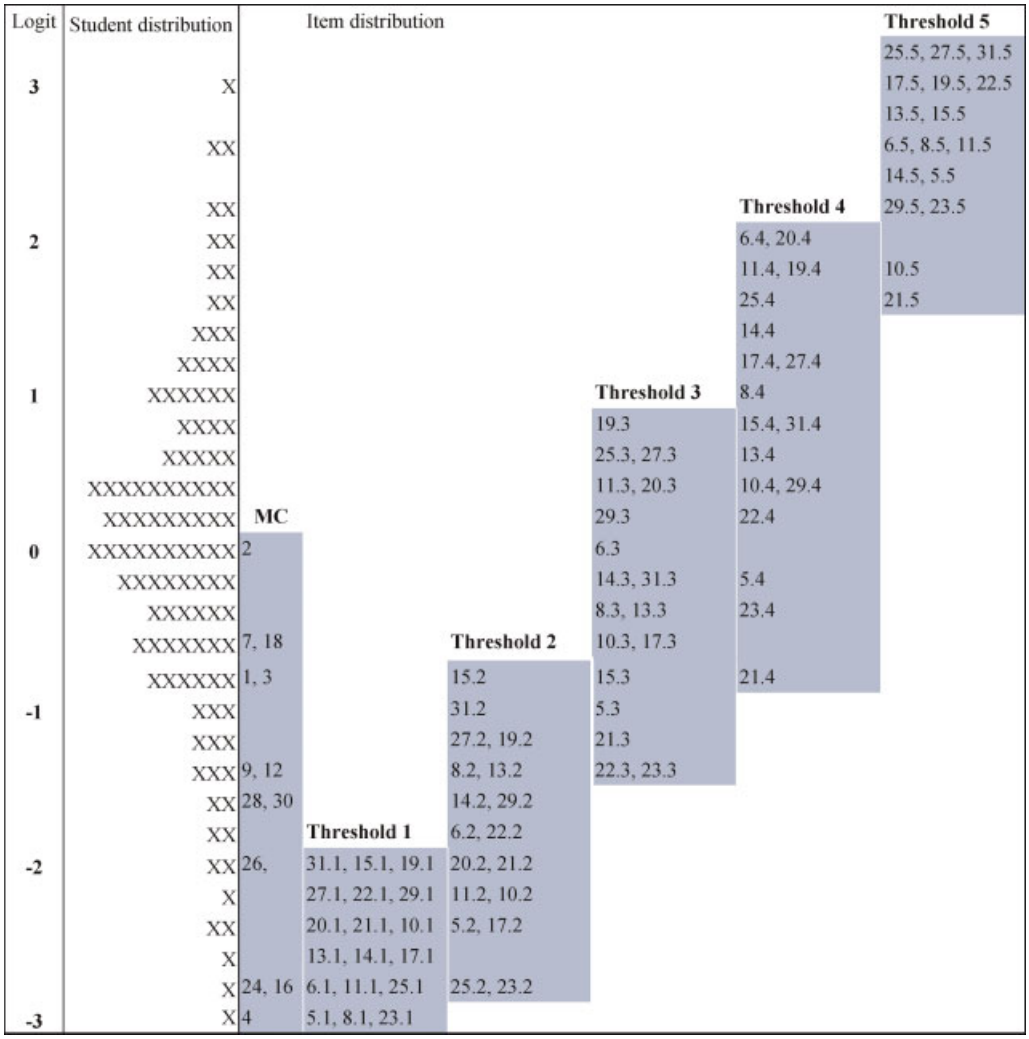

Figure 2. Distribution item difficulty and student ability estimates. MC, multiple-choice items; Thresholds 1-5 represent the difficulty of scoring $k + 1$ instead k ($k = 0, 1, \ldots 5$). For example, Threshold 1 is the difficulty of scoring 1 instead of zero, and Threshold 5 is the difficulty of scoring 5 instead of 4. Threshold 6.1 stands for Threshold 1 on item 6.

range of the outfit statistic is between 0.70 and 1.30 (Wright & Linacre, 1994). A smaller outfit value suggests that the item does not contribute to the test information beyond what is provided by other items. A larger outfit value suggests that the item may measure a different construct from the rest of the items. A larger outfit value is a more serious threat to the validity of the test than a smaller outfit value.

Results showed that the outfit statistic for most of the items (79 out of 86, 92%) tested in this study fell between the acceptable range of 0.70 and 1.30 (Figure 3). There are a couple of items showing large misfit on the high school chemistry and physics and middle school earth science and physical science modules. Examination of these misfit items revealed that most of the items were too difficult to effectively differentiate among students in the sample.

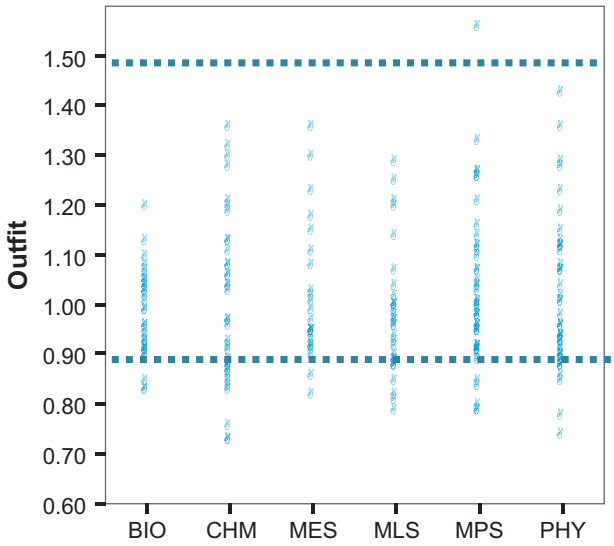

Figure 3. Outfit statistics. The acceptable range of outfit statistic is 0.70 and 1.30. The majority of the items fall in this range. MES, middle school earth science; MLS, middle school life science; MPS, middle school physical science; PHY, physics; CHM, chemistry; BIO, biology.

Person Reliability

Person reliability is calculated as $[1-(\mathrm{SEM}_i^2)/(\mathrm{var}(\theta)+\mathrm{SEM}_i^2)]$, where SEM_i stands for the SEM of ability estimate for student i, and $\mathrm{var}(\theta)$ stands for the variance of student ability estimates. Essentially, person reliability indicates the ratio of true ability variance to total variance. The majority of the person reliabilities fall between the reasonable range of 0.75 and 0.85 for all the six content subjects (Figure 4).

Gender

Males and females performed equally well on all of the six subjects, except for on high school biology and physics (Table 6). Females significantly outperformed males on biology ($p < 0.05$), and males scored significantly higher on high school physics ($p < 0.01$). Effect size indicated by Cohen's d (1988) showed that the magnitude of the gender difference on biology was negligible at -0.07. The effect size was small to medium at 0.32 on physics.

We speculated that students from one school may have contributed to the relatively large effect size on physics. In that school, very few female students had enrolled in the physics class, and they had demonstrated low physical science performance in the past. Analysis of students from only this school shows a large gender effect size of 0.74. If these students are excluded from the analysis, the gender effect size of the rest of the students drops to 0.24. There is clearly between-school variation in gender differences.

ELL Status

ELL students significantly underperformed as compared to others for five of the six subjects. Examination of the effect sizes showed three small to medium effect sizes: high school biology ($d = 0.29$), middle school life science ($d = 0.25$), and high school physics ($d = 0.46$). In middle school physical science, performance was equal (Table 7).

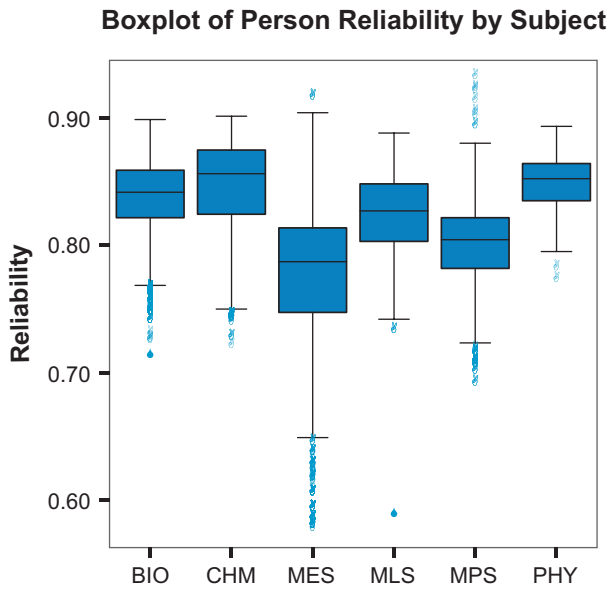

Figure 4. Boxplot of person reliability by subject.

Use of Computers for Homework

Students who reported using computers for homework showed superior performance for all six subjects (Table 8). The *p*-values from all the six *t*-tests are significant at the 0.001 level. Five of the six comparisons showed small to medium effect sizes, except high school chemistry ($d = 0.19$). Middle school physical science showed the largest magnitude of difference between the two groups of students, with an effect size of 0.50.

In order to further investigate whether student use of computers for homework was a mediator for students' socio-economic status (SES), we otained information on the percentage of students qualifying for free lunch and reduced-price lunch at each school and used such

Table 6
Gender comparison by module

	Male			Female				
	Mean	SD	*n*	Mean	SD	*n*	*t*	*d*
MES	0.00	0.59	2087	0.00	0.57	1997	−0.25	−0.01
MLS	−0.01	0.74	2106	0.03	0.71	2064	−1.59	−0.05
MPS	0.02	0.67	1910	−0.01	0.62	1751	1.48	0.05
PHY	0.11	0.77	335	−0.12	0.65	249	3.74**	0.32
CHM	0.01	0.81	879	−0.01	0.76	1088	0.37	0.02
BIO	−0.02	0.77	1968	0.03	0.74	2014	−2.29*	−0.07

MES, middle school earth science; MLS, middle school life science; MPS, middle school physical science; PHY, physics; CHM, chemistry; BIO, biology.

d = effect size calculated by dividing the mean difference between males and females by their pooled standard deviation.

*$p < 0.05$.

**$p < 0.01$.

Table 7
Comparison between ELL and non-ELL students

	Non-ELL			ELL				
	Mean	SD	n	Mean	SD	n	*t*	*d*
MES	0.02	0.57	3290	−0.08	0.60	802	4.32***	0.17
MLS	0.05	0.71	3217	−0.13	0.75	949	6.98***	0.25
MPS	0.01	0.64	2973	−0.02	0.65	673	1.15	0.05
PHY	0.11	0.75	393	−0.21	0.63	197	5.12***	0.46
CHH	0.03	0.77	1456	−0.07	0.82	528	2.50*	0.13
BIO	0.06	0.73	3072	−0.17	0.81	886	7.78***	0.29

ELL, English language learners; MES, middle school earth science; MLS, middle school life science; MPS, middle school physical science; PHY, physics; CHM, chemistry; BIO, biology.
d = effect size calculated by dividing the mean difference between ELLs and non-ELLs by their pooled standard deviation.
*$p < 0.05$.
***$p < 0.001$.

information as an indicator of the average SES of students at that school. We then calculated the correlation between the SES and the use of computers for homework at the school level. Results showed moderate correlations between the SES indicators and the use of computers for homework: $r = -0.57$ between percentage of free lunch and computer use, and $r = -0.42$ between percentage of reduced lunch and computer use. These results suggest that student use of computers for homework is more than a measure of their socio-economic status.

Online Versus Paper-and-Pencil Test Results

Students who responded to the online version performed significantly better than students who took the paper-and-pencil version across the six units (Table 9). The effect size of the difference ranges from 0.17 to 0.50, with the high school physics test showing the largest difference. The results suggest consistent advantages for the online testing mode.

Table 8
Comparison between students with different computer use status

	Use computers			Do not use computers				
	Mean	SD	*n*	Mean	SD	*n*	*t*	*d*
MES	0.11	0.56	2172	−0.12	0.57	1911	12.85***	0.40
MLS	0.12	0.68	2210	−0.11	0.75	1930	10.18***	0.32
MPS	0.13	0.61	2168	−0.18	0.65	1474	14.81***	0.50
PHY	0.09	0.74	396	−0.17	0.68	196	4.22***	0.37
CHM	0.04	0.78	1431	−0.10	0.79	546	3.75***	0.19
BIO	0.08	0.74	2438	−0.11	0.76	1518	7.91***	0.26

MES, middle school earth science; MLS, middle school life science; MPS, middle school physical science; PHY, physics; CHM, chemistry; BIO, biology.
d = effect size calculated by dividing the mean difference between computer users and non-users by their pooled standard deviation.
***$p < 0.001$.

Table 9
Comparison between students taking the online and paper/pencil tests

	Online			Paper/pencil				
	Mean	SD	*N*	Mean	SD	*N*	*t*	*d*
MES	0.08	0.54	2948	−0.02	0.66	1181	4.96***	0.17
MLS	0.16	0.64	1732	−0.11	0.77	2496	12.26***	0.36
MPS	0.16	0.57	1775	−0.15	0.67	1921	4.94***	0.40
PHY	0.24	0.69	267	−0.20	0.69	335	7.78***	0.50
CHM	0.19	0.60	431	−0.06	0.82	1585	5.75***	0.44
BIO	0.14	0.71	838	−0.04	0.76	3220	6.19***	0.29

MES, middle school earth science; MLS, middle school life science; MPS, middle school physical science; PHY, physics; CHM, chemistry; BIO, biology.
d = effect size calculated by dividing the mean difference between the online and paper/pencil groups by their pooled standard deviation.
***$p < 0.001$.

Impact of Multiple Variables

Multiple regression analysis was conducted to examine the impact of multiple variables including grade level, gender, ELL status, computer use for homework, delivery format, and number of math and science courses taken (Table 10). The purpose was to investigate the impact of each variable on student science performance with the presence of other variables. A separate multiple regression analysis was conducted for each of the six subjects. The outcomes variables are student knowledge integration proficiency indicated by ability estimates produced from ConQuest.

Results showed that grade was a significant predictor of performance on all subjects except for high school chemistry. It is expected that the higher the grade students are in, the more knowledge integration abilities they will have developed in learning science. After

Table 10
Summary results from multiple regressions by module

	Outcomes variable (student ability estimate)					
Predictors	MES	MLS	MPS	PHY	CHM	BIO
Grade	3.30***	2.78*	6.02***	2.17*	1.58	6.73***
Gender	−0.32	0.57	0.01	−3.49***	−0.85	1.52
English language learners	−1.90*	−2.17*	0.72	−2.80**	−1.39	−3.75**
Use computer for homework	12.30***	3.16**	2.17*	2.34*	1.57	4.83***
Delivery format	1.64	3.28**	1.99*	2.75**	3.83***	2.48**
Number of math courses taken	—	—	—	2.31*	6.80***	6.14***
Number of science courses taken	—	—	—	0.21	0.93	2.07*

The numbers represent the *t*-value of each predicator in the multiple regression for each module. Note that the multiple regression was run separately for each of the six modules. The variables number of math and science courses taken only apply to high school students. MES, middle school earth science; MLS, middle school life science; MPS, middle school physical science; PHY, physics; CHM, chemistry; BIO, biology.
*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

controlling for other variables, gender was a significant predictor for performance only on high school physics. ELL status was significantly associated with lower scores on high school biology and physics and middle school earth science and life science. Use of computers for homework was significantly associated with better performance on five of the six subjects, except high school chemistry. Delivery format was a significant predictor for all units except middle school earth science. Number of math courses taken significantly predicted performance on high school biology, physics, and chemistry. Number of science courses taken significantly predicted science achievement on high school biology.

Teachers' Implementation and Student Performance

In this analysis, we examined the relationship between the number of times the teachers had taught the inquiry-based units and the knowledge integration performance of their students. Altogether, between 2005 and 2008, 228 unique teachers implemented the inquiry-based units, including 160 one-time teachers, 42 two-time teachers, 19 three-time teachers, and 7 four-time teachers. A one-way analysis of variance was conducted using the number of implementations of the inquiry-based units as a factor and the mean student knowledge integration ability estimate (with possible range of -3 to 3) as the outcomes variable. Results revealed a significant performance difference among students taught by teachers with varying numbers of implementations of the units (Table 11). Compared to the students of teachers who taught the inquiry-based units one, two, and three times, students of teachers who taught the units four times had more success (Figure 5).

Discussion and Conclusions

This study reports on a 4-year effort to design and validate knowledge integration assessments of inquiry science learning. From 2005 to 2008, four different cohorts of students learned inquiry-based science instruction on traditionally hard-to-learn science topics. Knowledge integration assessments were designed to measure student understanding and explanation of science phenomena.

The psychometric analyses presented in this paper suggest that the assessments are able to adequately measure a range of student ability, including recognizing a correct scientific definition, generating an explanation to a science phenomenon, using evidence to justify argumentation, and establishing connections between different science ideas. A combination of assessment formats, including multiple-choice, short answer, and constructed-response items, contributes to effective measurement of knowledge integration abilities.

The Impact of Gender, ELL Status, and Computer Use

Among the three student background variables we investigated, gender, in general, showed the smallest achievement difference. Males and females performed equally on four out of six inquiry-learning topics. Females outperformed males on biology ($p < 0.05$), but the magnitude of the difference was negligible ($d = -0.07$). Males scored significantly higher

Table 11
ANOVA results using the number of times teaching TELS modules as the factor

	Sum of squares	df	Mean square	*F*	Sig.
Between groups	1.67	3	0.56	3.39	0.02
Within groups	36.77	224	0.16		
Total	38.44	227			

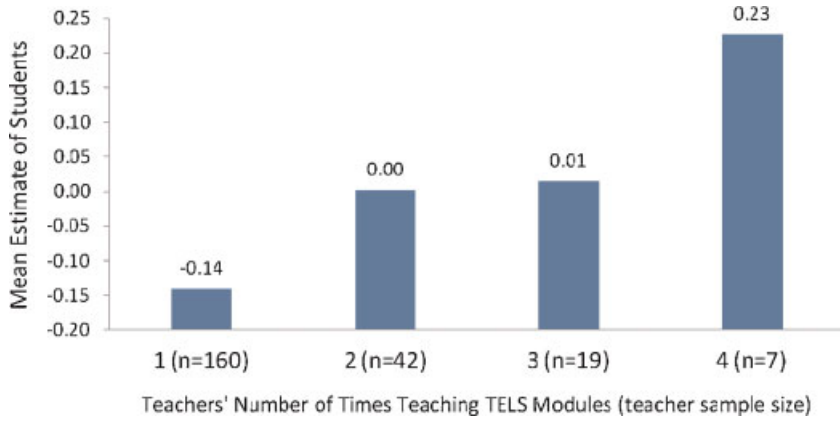

Figure 5. Mean student knowledge integration estimate by teachers.

on physics ($p < 0.01$), with a small to medium effect size of 0.32. Further analysis was conducted for one class that showed an unusual pattern of gender difference in physics due to a very small number of girls in the class. In general, we did not observe any prominent gender difference in the knowledge integration ability assessed in this study.

Studies have documented that girls may benefit in particular from inquiry-based learning (Burkam, Lee, & Smerdon, 1997; Cavallo & Laubach, 2001). The science instruction in this study featured hands-on activities and inquiry-learning opportunities for students to interact with science phenomena and manipulate variables through dynamic technology programs. The instruction activities also allow students to discuss ideas and communicate science findings in pairs. The authentic inquiry-learning experiences may contribute to the equal performance between males and females.

Compared to the small gender performance differences, differences with regard to ELL status and computer use were more substantial. ELL students were outperformed by non-ELL students on five of the six units. Among the five units that showed significant performance differences, three had effect sizes larger than 0.20. The physics unit showed the largest effect size of 0.46. As the percentage of ELL students increases in U.S. schools, in order to address the low academic achievement of ELL students, it is important to further understand how ELL students approach science learning. Currently many teachers are not sufficiently prepared to meet the special learning needs of ELL students. One of the key solutions is to enhance curriculum and teacher capacity in dealing with students with limited English language proficiency (Clark, Nelson, Atkinson, Ramirez-Marin, & Medina-Jerez, in press; Davis & Krajcik, 2005; Lee et al., 2009).

To help ELL students, it is important that teachers find ways to diagnose the source of the difficulty (such as lack of science content knowledge or insufficient English language proficiency) and find ways to improve classroom materials. Research shows that adapting materials for English learners can benefit students (Clark et al., in press; Sisk-Hilton, 2009). Clark et al. embedded a language support tool in their online science instruction system to help English language learners understand science content. The language support tool provides audio help in the students' native language and also offers ''pop ups'' to explain words in English that students may not understand. Students welcomed the use of the language support tool. Future curriculum design should consider the use of such educative materials in order to best help language learners.

Several studies have also shown that science instruction emphasizing hands-on, inquiry-based activities with personally relevant problems benefits diverse learners including ELL students (Lee et al., 2005). Hands-on activities provide ways to gain scientific understanding through observation, investigation, and communication. Incorporating visualization tools into science assessment has benefits for all students, including ELL students (Kopriva & Sexton, 1999). Inquiry science can help by requiring students to communicate science findings. When students talk about science, they improve their English proficiency, expand vocabulary, and develop scientific writing and speaking skills (Lee et al., 2005). Constructed-response science assessment provides one kind of communication opportunity for ELL students. Turkan and Liu (in preparation) found that, contrary to the commonly held belief that ELL students may be disadvantaged by constructed-response items because of their language deficit, the open-response item format may help ELL students express their science ideas freely in a less constrained context.

The advantage of using computers for homework, indicated by statistical significance, was consistent across all six instructional units in the study, with middle school physical science showing the largest effect size of 0.50. Previous research has shown that the effect of computer use on academic achievement varies depending on how computers are used, and also the context of the usage (Papanastasiou et al., 2003; Wenglinsky, 1998). Using computers for educational purposes has had a positive effect on education (Christmann & Badgett, 1999; James & Lamb, 2000; Taningco & Pachon, 2008). In this study, we specifically asked about students' use of computers for homework, instead of their general use of computers, to focus on academic uses of computers. Our further analysis also provided evidence that student use of computers was not solely a reflection of socio-economic status. The finding suggests that use of technology offers consistent advantages to students in their inquiry-science learning. As a next step in this line of research, it would be interesting to pinpoint the specific ways that student computer use for homework helps improve their science performance. For example, when learning physical science, do students take advantage of education software to observe unseen physical systems (e.g., particles) and phenomena (e.g., global warming)? Do students use visualizations to understand the physical processes of force and motion? Do students manipulate variables to observe the physical outcomes? Research addressing these questions will advance our understanding of the specific features of computer use that benefit student science learning.

Advantages of Online Testing

Our analysis of the test delivery format showed a consistent advantage for students who took the online version over those who took the paper-and-pencil version. A number of reasons may contribute to the higher performance of the online group. First, all of the student participants in the study learned all the science units in an online environment. They also practiced assessments online during daily instruction, either individually or in pairs. Therefore, students' familiarity with the online instruction materials and daily assessments may have helped them succeed on the online annual assessment. Second, since the annual assessment was generally administered within a class period, students were required to complete the test within a time limit. Given the number of constructed-response items in the assessment, the online format may have allowed for greater efficiency when typing and revising responses. Third, we speculate that the students may have been more motivated to take the online assessment, since some of the items were presented in colorful graphs and charts, while everything in the paper-and-pencil test was black and white.

The strong evidence of the advantages of online tests raises important implications for future research and also for test developers in science education. When designing inquiry-based science assessments, we should consider using online testing more often to take advantage of its easy administration, easy data retrieval, support for writing, and its potential advantage in engaging and motivating students.

Item Format and Scoring Rubrics

In this study, we used a combination of multiple-choice, short-answer, and constructed-response items to measure knowledge integration. About half of the constructed-response items were stand-alone items, and half prompted the students to explain their response to a paired multiple-choice item. Analysis of the constructed-response items showed that they displayed satisfactory psychometric properties in addition to their apparent advantages in eliciting student reasoning and thinking. They fit the PCM well. They also had high reliabilities, which is usually not the case for a limited number of constructed-response items. We speculate that both the pairing with multiple-choice items and the use of differentiating scoring rubrics contributed to the efficiency of the constructed-response items used in this study. When a multiple-choice and a constructed-response item are paired, the preceding multiple-choice item may help focus the students' responses when they answer the constructed-response item. Both the stem and choices in the multiple-choice item may guide student thinking about their explanations in the constructed-response portion. Without the multiple-choice portion, some students may focus on less relevant or even off-topic ideas. Such responses will lower the reliability of the constructed-response items. In this sense, the multiple-choice part does a good job of scaffolding students by provided limited choices.

A carefully designed and differentiating scoring rubric also contributes to the effectiveness of the constructed-response items. The scoring rubrics developed for this study were aligned with the knowledge integration construct, in that the rubrics rewarded student ability in generating multiple science ideas and seeking connections between ideas in explaining science phenomena. The levels on each rubric represented a progression of student knowledge integration ability and were able to capture the many performance scenarios that lie between completely incorrect and completely correct. Findings from this study suggest that, when well designed, both multiple-choice items and constructed-response items can contribute to enhanced measurement of student inquiry-science ability (Liu et al., 2010b).

Recommendations for Designing Assessments That Capture Complex Thinking

Based on results from this study, we provide three recommendations for designing science assessment that measure complex thinking such as the knowledge integration assessments: (a) assessment should take advantage of multiple assessment formats for effective testing. Both multiple-choice and constructed-response items have their benefits and shortcomings, and utilizing both helps achieve a balance between in-depth assessment and testing efficiency. To improve the validity of the choices of multiple-choice items, pilot studies can be conducted to identify common student misconceptions and popular beliefs in constructing the distractors. When designing constructed-response items, be careful to situate the items in a context that is meaningful to the students. Another important feature of complex science assessment is that the assessment should offer ample opportunities for students to express their view of science phenomena. Elaborated student responses will provide valuable information for teachers to determine the understanding of their students and revamp instruction accordingly. (b) Assessment professionals may realize that the design of the scoring rubrics is as important as test development. The scoring rubrics should be designed to reflect the

intention of the assessments. Many times the good intention of constructed-response items to measure deep science understanding is impeded by superficial scoring rubrics that only recognize right or wrong answers. The misalignment between assessments with good face validity and ineffective scoring rubrics is one of the most serious caveats in measuring complex science reasoning. The design of scoring rubrics should be an iterative process starting with the notion to differentiate between students of varying understanding levels. The rubrics need to be distinctive enough to represent meaningful response categories but also exhaustive enough to capture all possible student answers. And (c) as more states consider the use of computer-based tests for effective assessment, researchers should take full advantage of online testing for its efficiency. Although not explored in this study, for low-stakes formative science assessment, online testing has the potential of providing instant feedback to students, as feedback has been proven to benefit student learning (Epstein, Epstein, & Brosvic, 2001; Epstein et al., 2002). Coupled with automated scoring techniques (Sukkarieh & Pulman, 2005), online testing can also allow teachers to make timely decisions about instruction based on assessment results.

Potential Teacher Impact and Next Steps of Research

Finally, we found that teachers who implemented the inquiry-based science units more often tended to have larger student success in science achievement. This finding raised a great interest in further exploring the relationship between teachers' curriculum implementation and student inquiry-science learning. Previous research has shown that many variables concerning teacher implementation may have a profound impact on student learning outcomes, including fidelity of teaching the knowledge integration modules (Lee et al., 2005), teachers' embracing of the inquiry-based pedagogy curriculum (Marx, Blumenfeld, & Krajcik, 1998; Schneider, Krajcik, & Blumenfeld, 2005), teachers' use of assessment (Davis & Krajcik, 2005; Ruiz-Primo & Furtak, 2006), and teachers' use of technology (Clark & Linn, 2003). In future research, we will focus on investigating how the different aspects of teachers' instruction affects student learning of inquiry science.

References

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. Educational Assessment, 8, 231–257.

Adedi, J., & Gandara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. Educational Measurement: Issues and practice, 26, 36–46.

Alonzo, A.C., & Steedle, J.T. (2009). Developing and assessing a force and motion learning progression. Science Education, 93, 389–421.

American Physical Society. (2007). Gender equity: Strengthening the physics enterprise in Universities and National Laboratories. Washington, DC: American Physical Society.

Baxter, G.P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. Educational Measurement: Issues and Practice, 17(3), 37–45.

Bodmann, S.M., & Robinson, D.H. (2004). Speed and performance differences among computer-based and paper/pencil tests. Journal of Educational Computing Research, 31, 51–60.

Bransford, J.D., Brown, A.L., & Cocking, R.R. (2000). How people learn: Brain, mind, experience, and school. Washington, DC: National Academy Press.

Briggs, D.C., Alonzo, A.C., Schwab, S., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. Educational Assessment, 11, 33–63.

Britner, S.L. (2008). Motivation in high school science students: A comparison of gender differences in life, physical, and earth science classes. Journal of Research in Science Teaching, 45(8), 955–970.

Brotman, J., & Moore, F.M. (2008). Girls and science: A review of four themes in the science educational literature. Journal of Research in Science Teaching, 45, 971–1002.

Buck, G., Mast, C., & Ehlers, N. (2005). Preparing teachers to create a mainstream science classroom conducive to the needs of English-language learners: A feminist action research project. Journal of Research in Science Teaching, 42(9), 1013–1031.

Burkam, D.T., Lee, V.E., & Smerdon, B.A. (1997). Gender and science learning early in high school: Subject matter and laboratory experiences. American Educational Research Journal, 34, 297–331.

Catsambis, S. (1995). Gender, race, ethnicity, and science education in the middle grades. Journal of Research in Science Teaching, 32, 243–257.

Cavallo, A.M.L., & Laubach, T.A. (2001). Students' science perceptions and enrollment decisions in different learning cycle classrooms. Journal of Research in Science Teaching, 38, 1029–1062.

Chi, M.T.H. (1998). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In: R. Glaser (Ed.), Advances in instructional psychology. Mahwah, NJ: Lawrence Erlbaum Associates.

Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5, 121–152.

Christmann, E., & Badgett, J. (1999). A comparative analysis of the effects of computer-assisted instruction on student achievement in differing science and demographical areas. Journals of Computers in Mathematics and Science Teaching, 18, 135–143.

Clark, D., & Linn, M.C. (2003). Designing for knowledge integration: The impact of instructional time. The Journal of the Learning Sciences, 12, 451–494.

Clark, D.B., Nelson, B., Atkinson, R., Ramirez-Marin, F., & Medina, W. (in press). Integrating flexible language supports within online science learning environments. In: H. Waxman (Ed.), Research on technology use in multicultural set.

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

Davis, E.A., & Krajcik, J. (2005). Designing educative curriculum materials to promote teacher learning. Educational Researcher, 34, 3–14.

De Jong, E., (2003). After exit: Achievement patterns of former English language learners. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

diSessa, A.A. (1993). Toward an epistemology of physics. Cognition and Instruction, 10(2/3), 105–225.

Duschl R. Schweingruber H. & Shouse A. (Eds.) (2007). Taking science to school: Learning and teaching science in grades K-8. Washington, DC: The National Academies Press.

Epstein, M.L., Epstein, B.B., & Brosvic, G.L. (2001). Immediate feedback during academic testing. Psychological Reports, 88, 889–894.

Epstein, M.L., Lazarus, A.D., Calvano, T.B., Matthews, K.A., Hendel, R.A., Epstein, B.B., & Brosvic, G.L. (2002). Immediate feedback assessment technique promotes learning and corrects inaccurate first responses. The Psychological Record, 52, 187–201.

Educational Testing Service. (2005). TOEFL computer-based and paper-based tests. Princeton, NJ: Author.

Fadigan, K.A., & Hammrich, P.L. (2004). A longitudinal study of the educational and career trajectories of female participants of an urban informal science education program. Journal of Research in Science Teaching, 41, 835–860.

Farenga, S.J., & Joyce, B.A. (1999). Intentions of young students to enroll in science courses in the future: An examination of gender differences. Science Education, 83, 55–75.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial-ethnic and gender groups. Journal of Educational Measurement, 39, 133–147.

Government Accountability Office. (2006). No Child Left Behind Act: Assistance from education could help states better measure progress of students with limited English proficiency. Washington, DC: Government Accountability Office.

Hazari, Z., Sonnert, G., & Sadler, P.M. (2010). Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study. Journal of Research in Science Teaching, 47(8), 978–1003.

James, R., & Lamb, C. (2000). Integrating science, mathematics, and technology in middle school technology-rich environments: A study of implementation and change. School Science and Mathematics, 100, 27–36.

Johnson, R.L., McDaniel, F., & Willeke, M.J. (2000). Using portfolios in program evaluation: An investigation of interrater reliability. American Journal of Evaluation, 21, 65–80.

Kim, J., & Herman, J.L. (2009). A three-state study of English learner progress. Educational Assessment, 14(3/4), 212–231.

Kingston, N.M. (2009). Comparability of computer- and paper-administered multiple-choice tests for k-12 populations: A synthesis. Applied Measurement in Education, 22(1), 22–37.

Kirsch, I., Braun, H., Yamamoto, K., & Sum, A. (2007). America's perfect storm: Three forces changing our nation's future. Princeton, NJ: Educational Testing Service.

Kolen, M.J., & Brennan, R.L. (2004). Test equating, scaling, and linking: Methods and practices. (2nd edn.). New York: Springer-Verlag.

Kopriva, R., & Sexton, U.M. (1999). Guide to scoring LEP student responses to open-ended science items. Washington, DC: Council of Chief State School Officers.

Krajcik, J.S., & Sutherland, L.M. (2010). Supporting students in developing literacy in science. Science, 328, 456–459.

Lee, H.S., Linn, M.C., Varma, K., & Liu, O.L. (2010). How do technology-enhanced inquiry science units impact classroom learning? Journal of Research in Science Teaching, 47, 71–90.

Lee, H.S., Liu, O.L., & Linn, M.C. (2011). Construct validity of inquiry assessments: Role of multiple choice and explanation item format. Applied Measurement in Education, 24, 115–136.

Lee, O., Deaktor, R.A., Hart, J.E., Cuevas, P., & Enders, C. (2005). An instructional intervention's impact on the science and literacy achievement of culturally and linguistically diverse elementary students. Journal of Research in Science Teaching, 42, 857–887.

Lee, O., Hart, J., Cuevas, P., & Enders, C. (2004). Professional development in inquiry-based science for elementary teachers of diverse student groups. Journal of Research in Science Teaching, 41, 1021–1043.

Lee, O., LeRoy, K., Thornton, C., Adamson, K., Maerten-Rivera, J., & Lewis, S. (2008). Teachers' perspectives on a professional development intervention to improve science instruction among English language learners. Journal of Science Teacher Education, 19, 41–67.

Lee, O., Maerten-Rivera, J., Penfield, R., LeRoy, K., & Secada, W.G. (2008). Science achievement of English language learners in urban elementary schools: Results of a first-year professional development intervention. Journal of Research in Science Teaching, 45, 31–52.

Lee, O., Penfield, R., & Maerten-Rivera, J. (2009). Effects of fidelity of implementation on science achievement gains among English language learners. Journal of Research in Science Teaching, 46, 836–859.

Linn M.C. (1995). Designing computer learning environments for engineering and computer science: The Scaffolded Knowledge Integration framework. Journal of Science Education and Technology, 4, 103–126.

Linn, M.C. (2006). The knowledge integration perspective on learning and instruction. In: K. Sawyer (Ed.) The Cambridge handbook of the learning sciences (pp. 243–264). New York: Cambridge University Press.

Linn, M.C., Davis, E.A., & Bell, P. (2004). Internet environments for science education. Mahwah, NJ: Lawrence Erlbaum Associates.

Linn, M.C., & Eylon, B-S. (2006). Science education: Integrating views of learning and instruction. In: P. A. Alexander & P. H. Winne (Eds.) Handbook of educational psychology (pp. 511–544). Mahwah, NJ: Lawrence Erlbaum Associates.

Linn, M.C., & Hsi, S. (2000). Computers, teachers, peers: Science learning partners. Mahwah, NJ: Lawrence Earbaum Associates.

Linn, M.C., Lee, H.S., Tinker, R., Husic, F., & Chiu, J.L. (2006). Teaching and assessing knowledge integration in science. Science, 313, 1049–1050.

Liu, O.L., Lee, H.S., Hofstetter, C., & Linn, M.C. (2008). Assessing knowledge integration in science: Construct, measures, and evidence. Educational Assessment, 13, 1–23.

Liu, O.L., Lee, H.S., & Linn, M.C. (2010a). Evaluating inquiry-based science modules using a hierarchical linear model. Educational Assessment, 15, 69–86.

Liu, O.L., Lee, H.S., & Linn, M.C. (2010b). An investigation of teacher impact on student inquiry science performance using a hierarchical linear model. Journal of Research in Science Teaching, 47(7), 807–819.

Liu, O.L., Lee, H.S., & Linn, M.C. (2011). A comparison among multiple-choice, constructed-response and explanation multiple-choice items. Educational Assessment, 16, 164–184.

Marx, R.W., Blumenfeld, P.C., & Krajcik, J.S. (1998). New technologies for teacher professional development. Teaching and Teacher Education, 14, 33–52.

Masters, G.N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149–174.

McCarthy, C.B. (2005). Effects of thematic-based, hands-on science teaching versus a textbook approach for students with disabilities. Journal of Research in Science Teaching, 42, 245–263.

McNeill, K.L., Lizotte, D.J., Krajcik, J., & Marx, R.W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. Journal of the Learning Sciences, 15(2), 153–191.

National Center for Education Statistics. (2004). Early childhood longitudinal study: Kindergarten cohort. K-Fifth grade longitudinal data. Washington, DC: National Center for Education Statistics.

National Center for Education Statistics. (2006). The nation's report card: NAEP science 2005 (NCES 2006-466). Washington DC: National Center for Education Statistics.

National Research Council. (1996). The national science education standards. Washington, DC: National Academic Press.

National Science Board, (2008). Science and Engineering Indicators. National Science Foundation Report, 08-01, retrieved January 18, 2009, from website: http://www.nsf.gov/statistics/seind08/.

New York City Department of Education. (2009). Diverse learners on the road to success: The performance of New York City's English language learners. New York, NY: New York City Department of Education, Office of English Language Learners.

Organization for Economic Co-operation and Development. (2007). PISA 2006 Science Competencies for Tomorrow's World. Paris: Organization for Economic Co-operation and Development.

Papanastasiou, E. (2002). Factors that differentiate mathematics students in Cyprus, Hong Kong, and the USA. Educational Research and Evaluation, 8, 129–146.

Papanastasiou, E., Zembylas, M., & Vrasidas, C. (2003). Can computer use hurt science achievement? The USA results from PISA. Journal of Science Education and Technology, 12, 325–332.

Payán, R.M., & Nettles, M.T. (2006). Current state of english-language learners in the U.S. K-12 student population. Princeton, NJ: ETS.

Race to the Top. (2009). Accelerating College and Career Readiness in States—Standards and Assessments. Author. Retrieved on June 22 2011 from http://www.achieve.org/files/RTTT-StandardsandAssessments.pdf.

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests (Copenhagen, Danish Institute for Educational Research, expanded edition (1980) with foreword and afterword by B.D. Wright). Chicago: The University of Chicago Press.

Rockow, M. (2008). This isn't English class! Using writing as an assessment tool in science. Science Scope, 31, 22–26.

Ruiz-Primo, M.A., & Furtak, E.M. (2006). Informal formative assessment and scientific inquiry: Exploring teachers' practices and student learning. Educational Assessment, 11, 237–263.

Sadler, P.M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distracter-driven assessment instruments. Journal of Research in Science Teaching, 35, 265–296.

Sandoval, W.A. (2003). Conceptual and epistemic aspects of students' scientific explanations. The Journal of the Learning Sciences, 12(1), 5–51.

Schneider, R.M., Krajcik, J., & Blumenfeld, P.C. (2005). Enacting reform-based science materials: The range of teacher enactments in reform classrooms. Journal of Research in Science Teaching, 42, 283–312.

Segall, D.O., (1997). Equating the CAT-ASVAB. In: W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), Computerized adaptive testing: from inquiry to operation (pp. 181–198). Washington, DC: American Psychological Association.

Sisk-Hilton, S. (2009). Teaching and learning in public: Professional development through shared inquiry. NY: Teachers College Press.

Smith, C.L., Wiser, M., Anderson, C.W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic molecular theory. Focus Article. Measurement: Interdisciplinary Research and Perspectives, 14, 1–98.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. Journal of Research in Science Teaching, 38, 553–573.

Stern, L., & Ahlgren, A. (2002). Analysis of students' assessments in middle school curriculum materials: Aiming precisely at benchmarks and standards. Journal of Research in Science Teaching, 39, 889–910.

Sukkarieh, J.Z., & Pulman S. (2005). Information extraction and machine learning: Automarking short free-text responses for science questions. Proceedings of the 12th international conference on artificial intelligence in education, Amsterdam, The Netherlands.

Taningco, M.R., & Pachon, H. (2008). Computer use, parental expectations, and Latino academic achievement. Los Angeles, CA: Tomas Rivera Policy Institute.

Thissen, D., Wainer, H., & Wang, X.B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. Journal of Educational Measurement, 31, 113–123.

Trends in International Mathematics and Science Study. (2003). Science concepts and science items. Washington, DC: Trends in International Mathematics and Science Study.

Turkan, S., & Liu, O.L. (under review). Differential performance by ELLs on an inquiry-based science assessment. International Journal of Science Education.

Wang, W., & Wilson, M.R. (1996). Comparing multiple-choice-items and performance-based items using item response modeling. In: G. Engelhard & M. Wilson (Eds.), Objective measurement: Theory into practice (Vol. 3, pp. 167–194). Norwood, NJ: Ablex.

Wenglinsky, H. (1998). Does it compute? The relationship between educational technology and student achievement in mathematics. Princeton, NJ: ETS Policy Information Center-Research Division.

Wilson, M. (2005). Constructing measures: An item response modeling approach. Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, B.D., & Linacre, J.M. (1994). Reasonable mean-square fit values. Rasch Measurement Transactions, 8, 370.

Wright, B.D., & Masters, G.N. (1982). Rating scale analysis. Chicago, IL: Mesa Press.

Wu, M.L., Adams, R.J., Wilson, M.R., & Haldane, S.A. (2007). ACERConQuest Version 2: Generalised item response modelling software. Camberwell: Australian Council for Educational Research.