International Journal of Science Education
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tsed20

Measuring Knowledge Integration Learning of Energy Topics: A two-year longitudinal study

Ou Lydia Liu[a], Kihyun Ryoo[b], Marcia C. Linn[c], Elissa Sato[d] & Vanessa Svihla[e]

[a] Educational Testing Service, Princeton, NJ , USA

[b] School of Education, University of North Carolina, Chapel Hill, NC, USA

[c] Education in Mathematics, Science, and Technology, University of California, Berkeley, CA, USA

[d] Graduate Group in Science and Mathematics Education, University of California, Berkeley, CA, USA

[e] Organization, Information and Learning Sciences, University of New Mexico, Albuquerque, NM, USA

Published online: 23 Mar 2015.

To cite this article: Ou Lydia Liu, Kihyun Ryoo, Marcia C. Linn, Elissa Sato & Vanessa Svihla (2015): Measuring Knowledge Integration Learning of Energy Topics: A two-year longitudinal study, International Journal of Science Education, DOI: 10.1080/09500693.2015.1016470

To link to this article: http://dx.doi.org/10.1080/09500693.2015.1016470

International Journal of Science Education, 2015
http://dx.doi.org/10.1080/09500693.2015.1016470

Measuring Knowledge Integration Learning of Energy Topics: A two-year longitudinal study

Ou Lydia Liu[a*], Kihyun Ryoo[b], Marcia C. Linn[c], Elissa Sato[d] and Vanessa Svihla[e]

[a]*Educational Testing Service, Princeton, NJ, USA;* [b]*School of Education, University of North Carolina, Chapel Hill, NC, USA;* [c]*Education in Mathematics, Science, and Technology, University of California, Berkeley, CA, USA;* [d]*Graduate Group in Science and Mathematics Education, University of California, Berkeley, CA, USA;* [e]*Organization, Information and Learning Sciences, University of New Mexico, Albuquerque, NM, USA*

Although researchers call for inquiry learning in science, science assessments rarely capture the impact of inquiry instruction. This paper reports on the development and validation of assessments designed to measure middle-school students' progress in gaining integrated understanding of energy while studying an inquiry-oriented curriculum. The assessment development was guided by the knowledge integration framework. Over 2 years of implementation, more than 4,000 students from 4 schools participated in the study, including a cross-sectional and a longitudinal cohort. Results from item response modeling analyses revealed that: (a) the assessments demonstrated satisfactory psychometric properties in terms of reliability and validity; (b) both the cross-sectional and longitudinal cohorts made progress on integrating their understanding energy concepts; and (c) among many factors (e.g. gender, grade, school, and home language) associated with students' science performance, unit implementation was the strongest predictor.

Keywords: *embedded assessment; knowledge integration; Rasch; WISE*

To measure dynamic, longitudinal progress in understanding complex science concepts, assessments that capture how students use evidence to explain their reasoning

*Corresponding author. Educational Testing Service, 660 Rosedale Rd, Princeton, NJ 08541, USA. Email: lliu@ets.org

are needed. Many researchers and policy-makers criticize current assessments for measuring the recall of isolated science ideas rather than emphasizing the connections among ideas (Pellegrino, Chudowsky, & Glaser, 2001; Shepard, 2007; Songer, 2006). In addition, research often assesses students' learning of energy concepts at a single point in time rather than measuring students' dynamic progress over time. In this study, we explore the reliability and validity of assessments of energy concepts that are designed to capture progress across two years of instruction.

Energy concepts play a vital role in everyday decision-making such as choosing energy-efficient heating solutions, selecting containers to keep food safe for a picnic, or keeping plants healthy. Energy is central to many science topics including mechanics, physics, chemistry, astronomy, biology, and geology (American Association for the Advancement of Science [AAAS], 1993, 2007; Bransford, Brown, & Cocking, 2000). Energy is a major focus of the Next Generation Science Standards (NGSS Lead States, 2013) in that energy-related topics appear in the physical science disciplinary core ideas (e.g. PS3.A definitions of energy; PS3.B conservation of energy; and PS3.C relationship between energy and forces) and energy is a cross-cutting concept (e.g. energy and matter). Despite the importance of energy, few state, national, and international tests measure understanding of energy concepts across courses and contexts (National Assessment Governing Board, 2004; Schmidt, Raizen, Britton, Bianchi, & Wolfe, 1997). In addition, most measures of energy concepts focus on recall and do not align with the goals of inquiry instruction. Quality assessments are needed to capture student progress over time and to evaluate the effectiveness of innovative, inquiry-oriented curriculum materials.

This article reports on the development and validation of energy assessments in the context of a longitudinal study. The assessments build on prior research on knowledge integration (Davis, 2003, 2004; Davis & Krajcik, 2005) and were implemented using the Web-based Inquiry Science Environment (WISE; Linn & Hsi, 2000). The assessments were validated on a cross-sectional and a longitudinal cohort of students who studied energy-related instructional units in WISE over two years. Previous research has validated assessments designed to measure knowledge integration (e.g. Lee & Liu, 2010; Liu, Lee, Hoftstetter, & Linn, 2008; Liu, Lee, & Linn, 2011a, 2011b) but very few studies have examined the assessments' effectiveness in documenting longitudinal growth of student learning. This research investigates whether these assessments can be used by teachers to measure student science learning across years.

Review of Research on Energy Assessment

Energy concepts are difficult to teach and assess because many existing curriculum materials emphasize abstract, inaccessible ideas (Nordine, Krajcik, & Fortus, 2011). For instance, middle school textbooks often define energy as the 'ability to do work'. However, the concept of work is not typically introduced until high school physics courses because it depends on the understanding of calculus (the integral of force over distance).

Liu and McKeough (2005) investigated developmental sequences of energy understanding by analyzing students' responses to 27 items related to energy topics on the Third International Mathematics and Science Study (TIMSS). They categorized the items into energy concept development levels and investigated whether items at higher cognitive levels target students with greater maturation. They found that the mean item difficulty estimates increased with the expected level of cognitive demands of the items. Additionally, older students tended to perform better on more difficult energy items than younger students. Although the authors argued that the findings supported the hypothesized developmental sequence of energy concepts, this argument requires further investigation, as the study did not consider the amount of instruction students received.

Nordine et al. (2011) studied the impact of instruction on the learning of energy ideas. The researchers designed a comprehensive energy unit with multiple lessons for middle school students. Using a multifaceted assessment approach involving energy content and concept questionnaires and student interviews, they compared students who completed the comprehensive energy unit with older students who received typical instruction. They found that students who studied the energy unit gave interview responses that were more aligned with experts' understanding than that of the older students. Students who completed the comprehensive unit also performed better on a benchmark energy assessment than did students in higher grades with typical instruction. These results suggest that comprehensive energy instruction is better than typical instruction for helping students interpret everyday energy-related science phenomena. The results also suggest that design of science learning experiences to promote coherent understanding can help students achieve a coherent understanding of energy.

Other researchers have relied on learning progressions (e.g. Duschl, Schweingruber, & Shouse, 2007) when studying energy. For example, Neumann, Boone, Viering, and Fischer (2013) developed a learning progression of energy concept for middle school students and designed a multiple-choice assessment, referred to as the Energy Concept Assessment, to measure students' energy knowledge. Through empirical validation, the authors found that the items tend to be more difficult when the content involves a higher level of conceptions of energy. In another study, Lee and Liu (2010) measured progress on a learning progression of energy topics across physical, life, and earth science domains among middle school students. They used students' responses to published science items on energy sources, conservation, and transformations. Learning progressions represent the increasingly sophisticated understanding of a topic (National Research Council, 2007). A typical learning progression assessment system measures characteristics of thinking associated with each understanding level, with the thinking progressing from naïve ideas to expert opinions. Lee and Liu (2010) found that energy conservation items were more difficult than items on energy transformations and sources, supporting the notion of developmental progression rather than natural maturation. They also found that students who took physical sciences courses scored significantly higher than students who took life and earth science courses supporting the importance of instruction in progress along a learning progression.

These studies support the argument that students typically hold a fragmented repertoire of ideas about energy and that energy understanding is primarily mediated by instruction (e.g. Nordine et al., 2011). Both the relevance of the science contexts and the appropriateness of the curriculum sequence affect students' progress in learning energy ideas.

A Knowledge Integration Approach to Design Energy Assessments

To improve students' cumulative, integrated understanding of energy across science disciplines, we used the knowledge integration framework to design web-based curricula units and assessments in WISE (https://wise.berkeley.edu/; Kali, 2006; Linn & Eylon, 2011; Slotta & Linn, 2009; Williams, DeBarger, Montgomery, Zhou, & Tate, 2012). Using energy as a core idea, we created three general science units (Thermodynamics, Plate Tectonics, and Global Climate Change) for the sixth grade and two life science units (Photosynthesis and Cellular Respiration) for the seventh grade. See Appendix for a detailed description of the units in WISE. The units integrate multiple aspects of energy, such as energy sources, energy transformation, and energy transfer (see Figure 1).

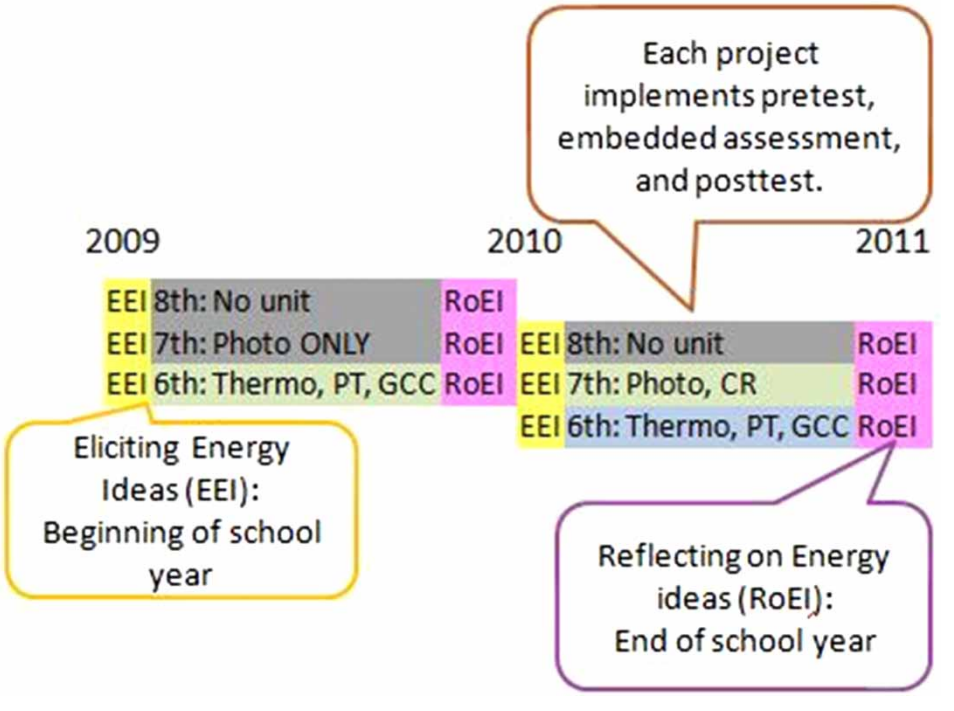

Figure 1. Assessment plan for measuring knowledge integration learning
Note: Photo = photosynthesis; Thermo = thermodynamics; PT = plate tectonics; GCC = global climate change; CR = cellular respiration

The units were designed to promote coherent understanding following the knowledge integration pattern that emphasizes eliciting, adding, distinguishing, and sorting out ideas (Linn, Davis,& Bell, 2004; Linn & Eylon, 2006). Using the knowledge integration framework, we designed assessments to assess students' ability to make connections among energy sources, energy transformation, and energy transfer across science topics. For instance, following the knowledge integration framework, units first elicit both the normative and non-normative ideas that students initially hold about the topic to encourage them to consider all of their ideas, including ideas reflecting their cultural and linguistic experiences as they encounter new views (Mayer, Dow, & Mayer, 2003; White & Gunstone, 1992).

Designing Assessments to Measure Integrated Understanding of Energy

To measure middle school students' understanding of energy, we designed items that assess how students build on the ideas they have learned in one course when they take another course. Conventional standardized tests often emphasize rote memorization or superficial learning rather than capturing students' cumulative understanding of science concepts. To assess progress in the understanding of core energy concepts, we designed an assessment system with distributed assessments: The beginning-of-year assessments to establish a baseline, pre-, and post-tests for each unit to measure immediate progress, embedded assessments during unit instruction to help with formative revisions, and end-of-year assessments to track student learning across units. This article focuses on the end-of-year assessments to provide an overall summary of students' understanding of energy after studying the units. Student scores on the beginning-of-year assessments were used to control for prior science learning.

Beginning- and End-of-Year Assessments

Following the knowledge integration framework, the beginning-of-year assessments were designed to elicit energy ideas and the end-of-year assessments were designed to encourage reflection on energy ideas (RoEI) while serving as a cumulative evaluation of students' understanding of the curriculum units (Figure 1). The beginning-of-year assessments were created from an item pool that had been designed and validated in prior National Science Foundation-funded research studies (Lee & Liu, 2010; Liu et al., 2008; Liu et al., 2011a, 2011b). The items demonstrated satisfactory internal consistency (e.g. Cronbach's alpha $>.70$), item fit, item difficulty, and coverage of students' ability range. The psychometric quality of the end-of-year assessment was examined in this study and reported in the Results section.[1]

The assessment development followed a rigorous design cycle in that prototype items were created, reviewed, and tested. Items that did not perform well were either removed or modified. Assessment reviewers included content experts, measurement experts, and science educators. Items were scrutinized for clarity, reading level, content coverage, and relevance to the instruction unit. In addition, various assessment considerations (e.g. use of the knowledge integration principles,

content coverage, varied item formats, scoring rubrics that reward coherent understanding) were embedded in the design process. These considerations made the assessments suitable for multiple grades in this longitudinal study. In addition, the empirical results reported in the Results section suggested that there was no ceiling or floor effect of the assessments for any of the grades tested in this study, and the psychometric properties of the assessments were satisfactory across grades.

Item formats. We employed a range of item types to measure understanding and to ensure that students had ample opportunity to demonstrate thinking and reasoning. Multiple-choice items, including two life science items, were selected from among the TIMSS published items. Other multiple-choice items were used in conjunction with constructed-response questions that asked students to explain their choices. We also used stand-alone constructed-response items and Energy Story (e.g. the Green Roof item where students construct a narrative about a science phenomenon involving energy). The beginning-of-year assessment had 11 items, including 6 multiple-choice, and 5 constructed-response items (i.e. explain your choice) that were coupled with multiple-choice items. The end-of-year assessment contained 20 items: it included 10 multiple-choice items, 9 constructed-response items associated with a multiple-choice item, and 1 Energy Story item.

Table 1. Assessment items, type, and energy concepts measured for the end-of-year assessment

Item	Type	Item origin	Energy sources	Energy transformation	Energy transfer
Items 1 & 2	MC + Exp	KI	Fire		Convection
Items 3 & 4	MC + Exp	KI			Conduction
Items 5 & 6	MC + Exp	KI	Sun or fire		Radiation
Items 7 & 8	MC + Exp	KI	Sun Food	Light energy→heat energy Heat energy→ IR Light energy→ chemical energy	
Items 9 & 10	MC + Exp	TIMSS 95 KI	Sun Food	Light energy→ chemical energy	Food chain
Items 11 & 12	MC + Exp	KI	Sun	Heat energy → IR	Energy balance
Items 13 & 14	MC + Exp	TIMSS 95 KI	Sun Food	Light energy → chemical energy	Food chain
Items 15 & 16	MC + Exp	KI	Core		Convection
Item 17	MC	KI			Conduction
Items 18 & 19	MC + Exp	KI			Conduction
Item 20	Energy Story	KI	Sun	Light energy→heat energy Heat energy→ IR Light energy→ chemical energy	Radiation

Note: MC = multiple-choice; Exp = explanation items; and KI = knowledge integration.

The use of multiple assessment formats ensured that the assessment did not primarily ask for isolated facts about one energy concept in a single science domain but rather engaged students in explaining how their ideas were connected. Thus, items involving constructed-responses required students to make scientifically valid links among normative energy ideas, such as energy sources, energy transformation, and energy transfer. The multiple-choice items measured fairly straightforward ideas yet were efficient as they typically elicit more information than do constructed-response items given the same test length (Wainer & Thissen, 1993), and thus contribute to the reliability of the overall assessments.

Content coverage. Items were carefully designed to align with the instructional units. Table 1 provides information about how each item on the end-of-year assessments targets the energy concepts emphasized in the curriculum units. Note that in the beginning-of-year assessment, although some items did not directly measure the core concepts (energy source, transfer, and transformation) covered in the units, they captured relevant energy concepts considered critical for students who are developing a complex understanding of energy (i.e. thermal equilibrium and the relationship between thermal conductivity and sensation; AAAS, 1993, 2007). The assessments were aligned with the curriculum units in that both aim to improve students' ability to integrate understanding across science ideas, domains, and contexts (Linn, Lee, Tinker, Husic, & Chiu, 2006).

To assess knowledge integration, the constructed-response items and Energy Story required students to make connections among different ideas about energy, rather than focusing on isolated, factual knowledge about a single energy concept. The assessments were designed to encourage students to explain their thinking and to refine their ideas. For example, an item called the Green Roof Energy Story[2] required students to write a coherent story about what happens to the energy from the sun when it hits a roof where plants are growing. To answer this question correctly, students must integrate their understanding of energy transformation and transfer learned in the sixth grade Global Climate Change unit with an understanding of how energy is transformed and stored during photosynthesis, learned in the seventh grade Photosynthesis unit.

Scoring rubrics. The multiple-choice items were scored dichotomously. The constructed-response items and Energy Stories were scored using a five-level rubric derived and customized from the knowledge integration rubric (Linn et al., 2006). The knowledge integration rubric emphasizes capturing the range of students' proficiency from irrelevant and non-normative ideas, to partial understanding of the connection between energy ideas, and to more elaborated links between those ideas. Table 2 shows an example of the customized knowledge integration rubric for the Green Roof item. A great deal of attention was paid to the levels in the individual rubrics to find ways to effectively differentiate among levels of integrated understanding. Note that the inter-rater reliability was over .90 in Pearson correlation for the constructed-response items, indicting good agreement among raters.

Table 2. Green roof item and knowledge integration scoring rubric

Scientifically valid links			
Link	Description		
Energy transformation	• Light energy from the sun is transformed into heat energy when it hits the roof • Light energy from the sun is absorbed by plants and transformed into chemical energy		
Energy source	• Chemical energy is stored in glucose (sugar and food) and used as an energy source for the plants to grow and function		
Energy transfer	• Energy from the sun moves through radiation, space, or wave		
Score	KI level	Description	Example
1	Off-task	No answer or off-task	'I don't really know'
2	No link	Non-normative or scientifically invalid links and ideas	'Energy comes from the plants.'
3	Partial link	Normative ideas without scientifically valid connections between ideas	'The energy comes from the sun and it will transfers to the plants because there grove the plant on their house. The energy will not go to the house because the plant cover the house head so the heat energy will not transfers.'
4	Full link	One scientifically valid and elaborated link between normative and relevant energy ideas	'The energy comes from the sun. It is transferred by the radiation. The energy is stored in the plant. It goes through the way a plant makes food.'
5	Complex link	Two scientifically valid links	'Light energy comes from the sun. The sun radiates it through space and over to Earth, where it enters the atmosphere and continues down towards the ground. When the light energy reaches the plants on top of the roof, it is absorbed by chloroplasts in the plants to combine with water and carbon dioxide and produce food and chemical energy for the plant and oxygen, which is released into the air.'

Objectives of this Study

This study served three main purposes: (a) to investigate the psychometric quality of the end-of-year energy assessment; (b) to examine the students' change in performance from Year 1 to Year 2 on the end-of-year assessment after controlling for prior ability, for both the cross-sectional and the longitudinal cohorts; and (c) to

investigate the factors that contributed to students' performance on the end-of-year assessment, after controlling for their prior science ability. In this investigation, we gathered information about students' grade, gender, language, school, and number of units studied as an indicator of energy-related inquiry instruction. We were particularly interested in examining the relationship between unit learning and performance as students varied in terms of the number of units that they learned. The following research questions were proposed in alignment with the purposes of this study:

(1) How valid, equitable, and reliable are the items for measuring energy understanding?
(2) How do the items differentiate students' performance from Year 1 to Year 2, after controlling for prior ability, for both the cross-sectional and the longitudinal comparisons?
(3) How are the various student characteristics (e.g. gender and language) and experiences (study of specific units) associated with assessment performance?

Methods

Design

At the start of Years 1 and 2, sixth-, seventh-, and eighth-grade science teachers administered the beginning-of-year test to elicit students' prior understanding of energy across science topics. During each academic year, the teachers implemented one or more of the units. Not all teachers implemented all available units for their grade level due to scheduling conflicts. At the end of each academic year, sixth-, seventh-, and eighth-grade science teachers administered the end-of-year assessment to

Table 3. Number of students by grade, gender, and language status

	Year 1 ($n = 2{,}037$)	Year 2 ($n = 2{,}310$)	Total ($N = 4{,}347$[a])
Grade			
6	514	733	1,247
7	971	912	1,883
8	552	665	1,217
Gender			
Male	862	1,075	1,937
Female	1,011	1,212	2,223
Missing	164	23	187
Language			
English	1,386	1,601	2,987
Other languages	487	686	1,173
Missing	164	23	187

[a]The total of 4,347 students includes 831 students who took the assessment in both years. Therefore, the unique total sample size is 3,516.

measure student learning over time. Administering the same assessments to students across grade levels ensured the comparability of student performance and allowed us to track longitudinal growth.

Participants

Participants were 4,347 students in grades 6–8 taught by 26 teachers in four middle schools in northern California. Note that the sample of 4,347 included a longitudinal cohort of 831 students who took the end-of-year assessments twice, first in 2010 (Year 1) and again in 2011 (Year 2). Therefore, the number of students was 3,516 in Years 1 and 2 combined.

See Table 3 for grade, gender, and language composition of the sample. The language variable had two categories: students who speak English at home, and students who speak a language other than English at home.

A total of 2,037 students took the end-of-year assessment in Year 1. The longitudinal cohort consisted of 831 of these students who also took the assessment in Year 2. The percentage of longitudinal record matching among all participants was about 41%. Students were excluded if they missed the Year 2 assessment, had over 30% missing data, or had teachers who did not continue with the project in Year 2. An additional 1,479 students were tested in Year 2, bringing the Year 2 total number of students to 2,310 (i.e. 831 + 1,479). Since the instruction targeted sixth and seventh graders, eighth graders were included in Year 1 to serve as a benchmark cohort for comparison with eighth graders in Year 2 who received instruction as seventh graders in Year 1. Therefore, by research design, the eighth graders in Year 1 did not take the assessment again.

The four participating middle schools represent a diverse student population in terms of socioeconomic status, home language, ethnicity, and Academic Performance Index scores. The participating teachers were 75% female. Most teachers had bachelor's degrees and more than five years of teaching experience. The teachers varied in terms of their attendance to the summer workshop offered by the research team: 55% attended both annual workshops, 30% attended one, and 15% attended none.

Analyses

In the following section, we describe the measurement methods used to estimate student ability, the reliability, and fit of the assessments. We also describe the analyses carried out to examine the (a) performance of gender and language groups, (b) students' cross-sectional and longitudinal performance change from Year 1 to Year 2, and (c) factors predicting students' performance.

Rasch-type models. The Rasch model (Rasch, 1960/1980) and Rasch partial credit model (PCM; Masters, 1982) were used to estimate student ability. The Rasch model was used for the dichotomously scored items and the Rasch PCM was used for the polytomously scored items. In Rasch-type models, the probability of a correct response to a given item depends on both the ability of the student and the difficulty level of the item.

The position of a student's ability relative to the item difficulty determines how likely a student is to provide the right answer to a given item. Rasch-type models use the logit scale to indicate students' ability estimates. A logit scale is a standardized interval scale. The space between each unit on the scale has a uniform value or meaning (Bond & Fox, 2001; Randall & Engelhard, 2009). The logit scale can range from negative infinity to infinity, but generally has a range of ± 3 (Harris, 1989). Each unit along the scale represents the amount of the underlying ability a student might have.

A student's raw score on a test is a sufficient statistic for the Rasch ability estimate. Students with the same raw score will obtain the same Rasch ability estimate. However, the intervals on the raw score scale may differ from the intervals on the Rasch logit scale (Boone & Scantlebury, 2006). For example, a one-point difference in raw scores may not correspond to a one-point difference in logits. In addition to the ability estimate, Rasch models also provide an estimate of the measurement error for each ability estimate, indicating the accuracy of the estimate. A major advantage of using Rasch-type models compared to the traditional raw score approach is that Rasch provides information on the accuracy of the estimation. Because of their mathematical simplicity and measurement strengths, Rasch-type models have been widely used in large-scale assessments such as the Programme of International Student Assessment (OECD, 2012). Increasingly, researchers draw on Rasch-type models when analyzing classroom-level assessments for formative purposes (Johnson & Tymms, 2011; Liu et al., 2008; Liu, Lee, & Linn, 2011a, 2011b; Rivet & Kastens, 2012; Siegel, 2007).

In this study, the software ConQuest was used to perform the Rasch and Rasch PCM analyses. ConQuest was able to adopt the models automatically for dichotomous and polytomous responses (Wu, Adams, Wilson, & Haldane, 2007). Students' ability was estimated using the marginal maximum likelihood estimation method in ConQuest and was scaled to be on the same scale as item difficulty estimates. The common values for students' ability estimates in logit numbers range from -3 to 3, with larger numbers indicating higher ability levels. In ConQuest, users can choose to constrain the estimation on the ability estimates or on the item difficulty estimates. If the former, the ability estimates would add up to zero; if the latter, the item difficulty estimates would add up to zero. In this study, the mean student ability estimates were used as the outcome variable for subgroup comparisons (e.g. gender, language, and year).

Person separation reliability. The PCM produces an estimate of reliability for each student. The person separation reliability indicates how efficiently a set of items is able to separate the students being measured, and is analogous to the Cronbach's alpha in classical test theory in that both are ratios of true measure variance to observed measure variance.

Item fit. Both the weighted and unweighted mean square fit statistics were examined. The weighted fit (infit) statistic detects abnormal patterns of responses when student ability and item difficulty are close, while the unweighted fit (outfit) statistic detects unexpected student responses on items that are either too difficult or too easy for them. Both statistics follow a chi-square distribution and can be transformed into a normal distribution with *t*-values larger than 1.96 suggesting misfit. A drawback with

this approach is that when the sample size is large, one can always expect large misfit values. A correction to this problem is to examine the actual fit statistics. A commonly used rule is that fit statistics between .70 and 1.30 correspond to a reasonable fit (Wright & Linacre, 1994), with values smaller than .70 suggesting redundancy in item content and values larger than 1.30 suggesting the possible presence of multiple constructs.

Cross-sectional and longitudinal comparisons. To compare the performance of the cross-sectional cohorts from Year 1 to Year 2, we need to control for their prior performance. An analysis of covariance (ANCOVA) following the general linear model feature in the statistical package SPSS was conducted for the cross-sectional cohorts, with student ability estimates on the end-of-year assessment as the outcome variable, year as a fixed factor, and students' performance on the beginning-of-year assessment as a covariate indicating students' prior performance. As mentioned earlier, 11 items were included in the beginning-of-year assessment to measure middle school energy concepts. Students' ability estimates from the beginning-of-year assessment using a Rasch PCM served as the control variable. Missing data were handled using pairwise deletion.

In addition to the ANCOVA, the predicted values of the ability estimates on the end-of-year assessment from the ANCOVA were used to compare the performance of the cross-sectional cohorts from Year 1 to Year 2. The reason that the predicted values were used was to control for prior performance. The comparisons were done separately for each grade and through independent sample *t*-tests. Results from the *t*-tests offer details about the statistical significance of the performance difference between the cross-sectional cohorts. Effect sizes indicated by standardized mean differences are also provided.

For the comparison of performance across two years for the longitudinal cohort, a similar ANCOVA was used. Paired sample *t*-tests were then used to compare the predicted values of ability estimates on the end-of-year assessment between Years 1 and 2 for each grade. For this cohort of students, we also examined the mean score for each item on the end-of-year assessment across two years. We used the Energy Story item to illustrate the progress students made from Year 1 to Year 2. We also compared, both cross-sectionally and longitudinally, the performance of students of different language status.

Impact of unit implementation on cumulative learning of energy. An ANCOVA using the general linear model in SPSS was conducted to investigate the relationship between unit implementation and student performance on the end-of-year assessment for all the students, with unit implementation and prior knowledge indicated by the beginning-of-year scores as covariates, and grade, gender, school, and language as fixed factors. In this case, the covariates were considered continuous variables and the fixed factors were categorical variables. The interaction terms were removed from the model as they were not of interest to this study. Partial eta squared values (η^2) were provided as an indicator of effect size for each independent variable. Note that although students were nested in schools, since there were only four schools involved, hierarchical linear modeling (HLM) with school as the second-level unit of analysis was not feasible. HLM was also not feasible using class/

teacher as the unit of analysis at the second level since; except for the school variable, all other variables were unique at the student level, including the unit of implementation.

Results

Reliability

The person separation reliability for the RoEI assessment with 20 items estimated from ConQuest was .73. The reliability was relatively lower at .60 for the 11-item beginning-of-year assessment. The lower reliability was probably due to the small number of items. However, it is still justified to include the beginning-of-year estimates as a control for prior ability.

Fit

All of the fit statistics fell between .70 and 1.30, suggesting reasonable fit. The correlation between the two types of fit statistics was .90. Item 17 showed the largest value of infit. Examination of this item revealed that it was one of the easiest items on the test, which may have contributed to its relatively large fit value.

Gender and Language Comparisons

Females performed statistically significantly better than males ($t = 2.24$, $p < .05$; Table 3), but the effect size indicated by Cohen's d (1988) was negligible ($d = .07$). Students who only speak English at home significantly outperformed students who speak a language other than English at home ($t = 1.99$, $p < .05$), although the effect size was small ($d = .15$). The performance difference observed in the language groups was smaller than what is reported in prior research (e.g. .30) using similar assessments administered in a paper and pencil format (e.g. Liu et al., 2011a, 2011b).

Cross-sectional Performance Change

The ANCOVA showed that both year and prior performance were significant predictors of end-of-year performance (year: $F_{df=1} = 55.6$, $p < .001$, $\eta^2 = .03$; prior performance: $F_{df=1} = 162.1$, $p < .001$, $\eta^2 = .10$). η^2 is the partial eta squared value and serves as an effect size. An η^2 value of .10 suggests that this independent variable explained about 10% of the variance in the outcome variable.

Table 4 shows that overall the cross-sectional cohorts made statistically significant progress on the end-of-year assessment from Year 1 to Year 2, after controlling for prior ability. The effect size of the performance difference was .13, indicating small but reasonable progress. Both sixth and seventh graders in Year 2 performed significantly higher than their respective counterparts in Year 1. Eighth graders performed similarly in Years 1 and 2. The effect size ($d = .16$) of the performance difference was

Table 4. Cross-sectional and longitudinal comparisons between Years 1 and 2

Cross-sectional	Year 1			Year 2				
Grade	*n*	Mean	SD	*n*	Mean	SD	*t*	*d*
All	1,206	−0.01	0.61	1,479	0.06	0.49	3.29**	0.13
Grade 6	137	−0.12	0.68	733	0.01	0.49	2.66**	0.22
Grade 7	517	0.00	0.62	535	0.12	0.51	3.43***	0.21
Grade 8	552	0.02	0.57	211	0.10	0.43	1.84	0.16

Longitudinal	Year 1				Year 2				
Grade	*n*	Mean	SD	Grade	*n*	Mean	SD	*t*	*d*
All	831	0.07	0.65	All	831	0.22	0.49	7.35***	0.26
Grade 6	377	0.07	0.68	Grade 7	377	0.28	0.49	6.88***	0.36
Grade 7	454	0.08	0.63	Grade 8	454	0.17	0.48	3.56***	0.17

**$p < .01$.

***$p < .001$.

smaller than that of the two other grades ($d = .22$ and .21 for sixth and seventh graders, respectively). In general, there is evidence that Year 2 students outperformed Year 1 students, after controlling for initial differences in their ability. The improvement may be attributed to the improved curriculum units and enhanced instruction from Year 1 to Year 2.

Longitudinal Performance Change

For the longitudinal cohort, both year and prior performance were significant predictors of end-of-year assessment scores (year: $F_{df=1} = 202.6$, $p < .001$, $\eta^2 = .12$; prior ability: $F_{df=1} = 200.8$, $p < .001$, $\eta^2 = .11$). Thus, both sixth and seventh graders made significant progress from Year 1 to Year 2 (Table 4). The seventh graders in Year 2 who completed the units in both their sixth and seventh grades experienced a larger gain (.36 standard deviations (SD)) than the eighth graders in Year 2 who had completed only one unit as seventh graders in Year 1 (.17 SD).

The longitudinal cohort also made progress on average on almost all the individual items from Year 1 to Year 2 (Figure 2). We use one item (Green Roof) to illustrate students' improvement from Year 1 to Year 2. The results on the Green Roof item demonstrated that students who received instruction for two years significantly improved their understanding of energy across science topics as compared to students who received instruction for only one year. The seventh-grade students who completed one seventh-grade unit in Year 1 and did not receive any instruction in Year 2 achieved a knowledge integration score of 3.07 (see scoring rubric in Table 2) on this item in Year 1. This indicates that overall, the seventh-grade students developed normative ideas about energy concepts. In particular, many students were able to identify the sun as the main source of energy and to explain that energy from the sun is used to help plants grow or make food. Although these students did not complete any other units during eighth grade, they scored moderately higher on the same item in Year 2 (mean score 3.38; $d = .35$) than in year 1.

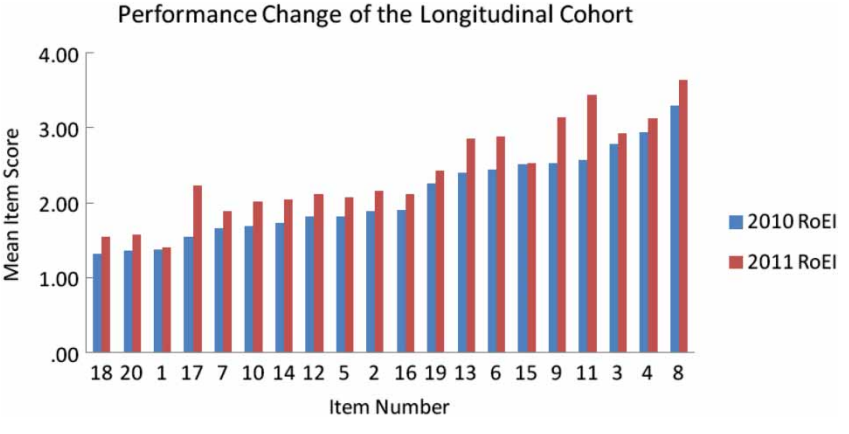

Figure 2. Performance change of the longitudinal cohort

The students who began instruction in sixth grade and received instruction for two years made the largest gain. These students' average score on the Green Roof item in Year 1 was 2.87, which suggests that overall the students had non-normative or irrelevant ideas about energy. However, after completing two more units during Year 2, their average score on the Green Roof item increased about one knowledge integration level to 3.42 ($d = .60$).

Progress by Language Status

For both the cross-sectional and longitudinal cohorts, students whose primary language was English improved very slightly more than students who speak another language at home (Figure 3). The difference was negligible (i.e. $d = .03$) in both cases. These findings confirm that language experience does not appear to hinder progress in integrated understanding resulting from study of WISE energy units. They also suggest that the small performance advantage for students who speak only English at home was not due to possible biases existing in items, but was more likely a reflection of the true achievement similarities between the two groups of students. The comparable gains for these groups suggest that the units were effective in serving diverse learners.

Relationship between Number of Units Studied and Learning Outcomes

Table 5 presents the results investigating the impact of unit learning on energy performance. After controlling for prior science knowledge, grade, gender, school, and language, the number of units studied was a significant predictor of science performance (Table 5). Number of units completed also had the largest effect size ($\eta^2 = .13$).

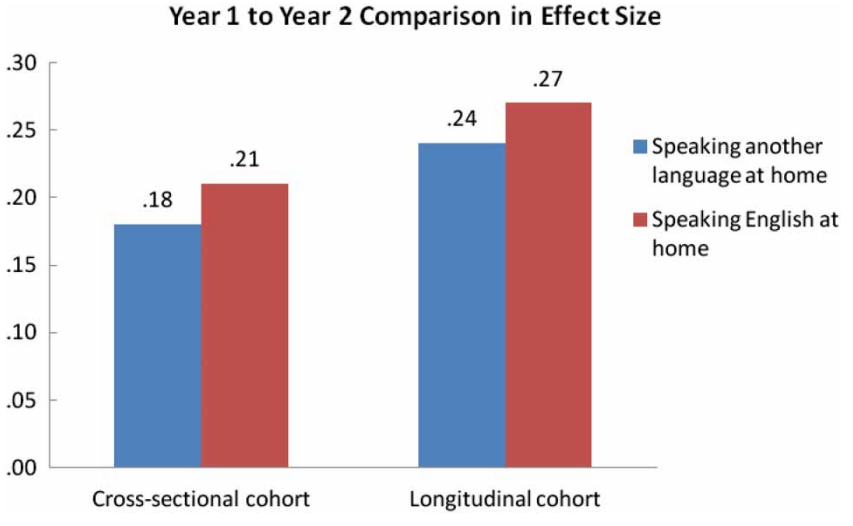

Figure 3. Comparisons by language status. $^*p < .05$

Table 5. Regression results

Source	Type III sum of squares	df	Mean square	F	η^2
Intercept	22.65	1	22.65	81.10***	.02
Prior science knowledge	131.52	1	131.52	467.86***	.11
Grade	6.02	1	6.02	21.27***	.01
Gender	.95	1	.95	3.00	.02
Language	2.59	1	2.59	9.15**	.01
School	42.95	3	14.32	50.58***	.03
Curriculum units	180.00	4	45.00	158.97***	.13
Error	1,045.36	3,693	.28		
Total	1,432.73	3,703			

**$p < .01$.

***$p < .001$.

After controlling for other factors, 13% of the variance in test scores can be explained by opportunity to learn.

Discussion

In this study, we designed assessments to measure students' progress in learning energy concepts from WISE inquiry units over time. The assessments were validated using two years of data from over 4,000 students in four middle schools.

A limitation of this study was that no control group was available given the long time span of this study. Therefore, all the results are correlational and no causal conclusion can be made between students' progress on science learning and the factors associated with students' learning (e.g. unit implementation). Nevertheless, the correlational results yielded some findings that are worth noting. We found that students made consistent improvement in the two-year period and that unit implementation was the strongest predictor of students' performance gains. The lack of a control group does not detract from conclusions about the psychometric quality of the assessments. The items overall demonstrated satisfactory psychometric properties. Evidence of learning gains associated with instructional experience supports the validity of the assessments. In the following section, we discuss the factors that may be associated with the non-native English speakers' learning gains and with the assessments' instructional sensitivity. We explore implications for future assessment design.

Advantages for Linguistically Diverse Learners

It is notable that students whose home language is not English made improvements comparable to students, whose home language is English, a finding in contrast to those of other studies (Lee, Maerten-Rivera, Penfield, LeRoy, & Secada, 2008; Lee, Penfield, & Maerten-Rivera, 2009). The standardized mean difference in terms of progress from Year 1 to Year 2 was .03 between students who speak

English and those who speak another language at home for both the cross-sectional and the longitudinal cohorts. In other studies, the standardized mean differences have ranged from about .26 to .33 between non-English-speakers and other students in terms of gain scores after a year-long intervention (Lee et al., 2008, 2009).

We speculate that the inquiry instruction and assessment platform in WISE may have contributed to the success of diverse learners in understanding energy across science topics. Learning science can pose challenges to linguistically diverse students because scientific language has many features that are different from students' everyday language (Fang & Schleppegrell, 2008; Halliday & Martin, 1993; Lee, 2005, 2008). Culturally and linguistically diverse learners may benefit from the sorts of supports found in knowledge integration instruction and assessments (Brown & Ryoo, 2008; Clark & Linn, 2003; Lee, 2005; Schleppegrell, 2004; Turkan & Liu, 2012). For example, both the units and the assessments feature dynamic visualizations that depict invisible concepts of energy and link this information to other science content. These visualizations may reduce the complexity of the energy topic and help students develop a coherent understanding of the content in the units (Buxton, 1999; Dixon, 1995; Lee, 2005; Ryoo & Linn, 2012). In studying the units, students were required to participate in scientific inquiry through extensive reading, writing, and speaking, giving all students experience in using scientific language. In the units, students had multiple opportunities to construct their own explanations using evidence, engage in arguments with peers, and refine their understanding by iteratively revising their responses. This experience appears to have benefitted both students who speak English at home and those who speak other languages at home.

In future research, we plan to include a control group of English language learners and gather further evidence of how the instruction and assessment features contribute to the science learning of students who speak English as a second language.

Instructional Sensitivity of Assessments and Implications for Future Design

During the development of the assessments, close attention was paid to the alignment between the energy concepts covered in the assessments and those emphasized in the instructional units, as the assessments were intended to provide measures of the degree of understanding students reached after receiving the instruction. Empirical evidence from multiple sources points to the instructional sensitivity of the assessments: (a) in the cross-sectional cohort comparisons, the second-year cohort outperformed the first-year cohort, (b) the longitudinal cohort showed score gains from Year 1 to Year 2, and (c) among an array of factors, studying WISE inquiry energy units was the strongest predictor of scores on the end-of-year assessments. In the cross-sectional comparison (Table 4), the second-year cohorts performed better than the first-year cohorts at each grade level. A number of factors may have contributed to the improved performance of the second-year cohort, including iterative refinement of the curriculum materials from Year 1 to Year 2, more customized support provided to teachers by the research team, teachers' greater familiarity with the instructional and assessment

materials, and teachers' improved pedagogical practices resulting from their experience in the first year.

In the longitudinal comparison, the seventh graders in Year 2 showed a larger gain (.36 SD) than the eighth graders in Year 2 (.17 SD, see Table 4). The amount of curriculum experience could possibly explain the difference in longitudinal improvement: by the end of Year 2, the seventh graders had studied the WISE units for two consecutive years while the eighth graders had studied the units for only one year. Another notable finding from Table 4 is that in Year 2, the seventh graders actually scored significantly higher than the eighth graders ($\Delta M = .11$, $t = 3.25$, $p = .001$, $d = .23$).

These findings also support the notion that students' gains in understanding energy are largely determined by relevant science instruction and experience rather than by natural maturation (Braun, Coley, Jia, & Trapani, 2009; Nordine et al., 2011). As Table 5 shows, after controlling for background variables and prior ability, units studied were the largest contributor to students' assessment performance.

As for designing energy assessments, this study provides a good example of how assessments can be aligned with instruction and serves as valid measures of student learning. Designing assessments along with the design of innovative learning materials is essential to ensure that the assessments measure the intended outcomes of the instruction. Many standardized tests are disconnected with new instruction approaches and may not provide valid information about the effectiveness of the new curricula. Co-design of assessments and curricula leads to outcome measures that provide a meaningful evaluation of the effectiveness of the curricula.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes

1. The assessments can be accessed at this website http://wise.berkeley.edu/webapp/preview.html?projectId=6525
2. The Green Roof item:
   Brent and Emilio heard that growing plants on the roof could lower energy usage. Write an Energy Story to explain to them what happens to energy from the sun in the picture. Remember to include:

- where energy comes from;
- how energy moves/transfers from place to place;
- where energy goes or is stored;
- how energy changes/transforms.

References

American Association for the Advancement of Science (AAAS). (1993). *Benchmarks for science literacy, Project 2061*. New York: Oxford University Press.

American Association for the Advancement of Science (AAAS). (2007). *Atlas of science literacy: Project 2061* (Vol. 2). Washington, DC: American Association for the Advancement of Science and the National Science Teachers Association.

Bond, T., & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.

Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education*, *90*(2), 253–269.

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

Braun, H., Coley, R., Jia, Y., & Trapani, C. (2009). *Exploring what works in science instruction: A look at the eighth-grade science classroom* (ETS Policy Information Report). Princeton, NJ: Educational Testing Service.

Brown, B. A., & Ryoo, K. (2008). Teaching science as a language: A 'content-first' approach to science teaching. *Journal of Research in Science Teaching*, *45*(5), 529–553.

Buxton, C. (1999). Designing a model-based methodology for science instruction: Lessons from a bilingual classroom. *Bilingual Research Journal*, *23*(2–3), 147–177.

Clark, D. B., & Linn, M. C. (2003). Scaffolding knowledge integration through curricular depth. *Journal of Learning Sciences*, *12*(4), 451–493.

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

Davis, E. A. (2003). Prompting middle school science students for productive reflection: Generic and directed prompts. *The Journal of Learning Sciences*, *12*(1), 91–142.

Davis, E. A. (2004). Knowledge integration in science teaching: Analyzing teachers' knowledge development. *Research in Science Education*, *34*(1), 21–53.

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

Dixon, J. K. (1995). Limited English proficiency and spatial visualization in middle school students' construction of the concepts of reflection and rotation. *Bilingual Research Journal*, *19*(2), 221–247.

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

Fang, Z., & Schleppegrell, M. J. (2008). *Reading in secondary content areas: A language-based pedagogy*. Ann Arbor: University of Michigan Press.

Halliday, M. A. K., & Martin, J. R. (1993). *Writing science: Literacy and discursive power*. London: Falmer Press.

Harris, D. (1989). An NCME instructional module on: Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice*, *8*(1), 35–41.

Johnson, P., & Tymms, P. (2011). The emergence of a learning progression in middle school chemistry. *Journal of Research in Science Teaching*, *48*, 849–877.

Kali, Y. (2006). Collaborative knowledge-building using the Design Principles Database. *International Journal of Computer Support for Collaborative Learning*, *1*(2), 187–201.

Lee, C. (2008). The centrality of culture to the scientific study of learning and development: How an ecological framework in education research facilitates civic responsibility. *Educational Researcher*, *37*(5), 267–279.

Lee, H. S., & Liu, O. L. (2010). Assessing learning progression of energy concepts across middle school grades: The knowledge integration perspective. *Science Education*, *94*(4), 665–688.

Lee, O. (2005). Science education and English language learners: Synthesis and research agenda. *Review of Educational Research*, *75*(4), 491–530.

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., 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* (2nd ed., pp. 511–544). Mahwah, NJ: Lawrence Erlbaum Associates.

Linn, M. C., & Eylon, B.-S. (2011). *Science learning and instruction: Taking advantage of technology to promote knowledge integration*. New York, NY: Routledge.

Linn, M. C., & Hsi, S. (2000). *Computers, teachers, peers: Science learning partners*. Mahwah, NJ: Lawrence Erlbaum 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., Hoftstetter, C. & Linn, M. C. (2008). Assessing knowledge integration in science: Construct, measures, and evidence. *Educational Assessment*, *13*, 33–55.

Liu, O. L., Lee, H. S., & Linn, M. C. (2011a). Measuring knowledge integration: Validation of four-year assessments. *Journal of Research in Science Teaching*, *48*(9), 1079–1107.

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

Liu, X., & McKeough, A. (2005). Developmental growth in students' concept of energy: Analysis of selected items from the TIMSS database. *Journal of Research in Science Teaching*, *42*(5), 493–517.

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

Mayer, R., Dow, G., & Mayer, S. (2003). Multimedia learning in interactive self-explaining environment: What works in the design of agent-based microworlds? *Journal of Educational Psychology*, *95*(4), 806–812.

National Assessment Governing Board. (2004). *Twelfth grade student achievement in America: A new vision for NAEP*. Washington, DC: Author.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grade K-8*. Washington, DC: The National Academies Press.

Neumann, K., Boone, W., Viering, T., & Fischer, H. E. (2013). Towards a learning progression of energy. *Journal of Research in Science Teaching*, *50*(2), 162–188.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.

Nordine, J., Krajcik, J., & Fortus, D. (2011). Transforming energy instruction in middle school to support integrated understanding and future learning. *Science Education*, *95*(4), 670–699.

OECD. (2012). *PISA 2009 technical report*. PISA, OECD publishing. Retrieved from http://dx.doi.org/10.1787/9789264167872-en

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Randall, J., & Engelhard, G. (2009). Differences between teachers' grading practices in elementary and middle schools. *The Journal of Educational Research*, *102*, 175–186.

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, IL: University of Chicago Press.

Rivet, A. E., & Kastens, K. A. (2012). Developing a construct-based assessment to examine students' analogical reasoning around physical models in earth science. *Journal of Research in Science Teaching*, *49*, 713–743.

Ryoo, K., & Linn, M. C. (2012). Can dynamic visualizations improve middle school students' understanding of energy in photosynthesis? *Journal of Research in Science Teaching*, *49*(2), 218–243.

Schleppegrell, M. (2004). *The language of schooling: A functional linguistics perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Schmidt, W. H., Raizen, S. A., Britton, E. D., Bianchi, L. J., & Wolfe, R. G. (1997). *Many visions, many aims (TIMSS volume 2): A cross-national investigation of curricular intentions in school science*. London: Kluwer.

Shepard, L. A. (2007). Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 279–303). Mahwah, NJ: Erlbaum.

Siegel, M. A. (2007). Striving for equitable classroom assessments for linguistic minorities: Strategies for and effects of revising life science items. *Journal of Research in Science Teaching*, *44*, 864–881.

Slotta, J. D., & Linn, M. C. (2009). *WISE science*. New York, NY: Teachers College Press.

Songer, N. B. (2006). BioKIDS: An animated conversation on the development of complex reasoning in science. In R. Keith Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 355–369). New York, NY: Cambridge University Press.

Turkan, S., & Liu, O. L. (2012). Differential performance by ELLs on an inquiry-based science assessment. *International Journal of Science Education*, *12*, 1–27.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*, 103–118.

White, R., & Gunstone, R. (1992). *Probing understanding*. London: The Falmer Press.

Williams, M., DeBarger, A. H., Montgomery, B. L., Zhou, X., & Tate, E. (2012). Exploring middle school students' conceptions of the relationship between genetic inheritance and cell division. *Science Education*, *96*(1), 78–103.

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

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

Appendix

WISE Inquiry Energy Units

General science units for the sixth grade: Thermodynamics, Plate Tectonics, and Global Climate Change. The sixth-grade units were built upon tested WISE units by incorporating new visualizations and inquiry-based activities to strengthen the students' understanding of the underlying energy concepts. The *Thermodynamics unit* was modified in order to elicit and develop student understanding of conduction as a means of energy transfer. Visualizations were added to help students explore thermal equilibration and conduction. Although energy transfer by radiation is not introduced in this lesson, there are opportunities for students to discuss it, as heat sources include the sun and a gas stove. The heat energy idea introduced in the Thermodynamics unit was subsequently developed in the Plate Tectonics unit.

The *Plate Tectonics unit* supported students in developing an integrated understanding of surface and sub-surface geologic processes with a focus on convection, a macroscopic process of heat energy transfer driven by changes at the molecular level, with conduction being an additional means of energy transfer within the earth. Students

explored the phenomenon of convection at both macro- and microscopic levels of representation and then investigated how convection relates to plate boundary types.

In the *Global Climate Change unit*, we designed a series of predict–observe–explain activities, in which observations were made via interactive visualizations that allowed students to observe energy transformations. Understanding this series of transformations is crucial to forming a mechanistic understanding of climate change, because it is the infrared radiation, not the solar radiation, that is reflected back toward the earth by greenhouse gases, increasing the global temperature. Thus, though similar visualizations had been used in the previous curriculum units, the prompts surrounding these units focused students' attention on the energy. This unit revisited energy transfer by conduction and also used visualizations to help students differentiate between the previously studied types of transfer involving heat energy with the newly introduced ideas of energy transfer by radiation/light.

Life science units for the seventh grade: Photosynthesis and Cellular Respiration. The *Photosynthesis* and *Cellular Respiration units* were designed to help the seventh-grade students build on energy ideas learned in the sixth grade and promote their cumulative understanding of energy flow in life science. The Photosynthesis unit supported students in linking the concept of energy transformation from solar radiation to heat energy, introduced in the Global Climate Change unit, to the process of energy transformation from solar radiation to chemical energy in the plant's cell. Interactive, dynamic visualizations in the Photosynthesis unit helped students explore how light energy is converted into chemical energy and how chemical energy is stored in glucose inside the cell (for more details, see Ryoo & Linn, 2012).

Building on energy ideas introduced in the Photosynthesis unit, the *Cellular Respiration* unit emphasized how energy stored in glucose is released and transferred. Students have multiple opportunities to make connections between the transformation of matter and energy to observable plant growth by conducting virtual experiments in the Cellular Respiration unit. For example, students performed a virtual experiment and investigated how the total amount of glucose made, used, and stored was changing depending on the presence of light.