Validating an Assessment for Tracking Students' Growth in Understanding of Energy from Elementary School to High School

Joseph Hardcastle, American Association for the Advancement of Science;

Cari F. Herrmann-Abell, American Association for the Advancement of Science;

George E. DeBoer, American Association for the Advancement of Science

Paper presented at the 2017 NARST Annual International Conference

San Antonio, TX

April, 25, 2017

Introduction

Energy is a central concept in science, cutting across multiple fields including physics, chemistry, biology, and engineering. Discussions about energy are also common in modern social discourse, including topics such as alternative sources of energy and renewable energy technologies. The importance of energy is evident in its being both a core idea and a cross-cutting concept in the *Next Generation Science Standards* (NGSS Lead States, 2013). The increased emphasis on energy in NGSS, and ultimately in the science curriculum, has resulted in a need for new assessments to measure students' knowledge of energy and for teachers and researchers to be able to track students' growth of understanding about energy.

Several research-based energy assessments are available; however, their usefulness to K-12 teachers is limited. The Energy Concept Assessment (Ding, 2007) and the Energy and Momentum Conceptual Survey (Singh & Rosengrant, 2003) are both popular energy assessments for algebra or calculus-based physics courses. These assessments are powerful tools for high-school or university teachers, but are too difficult for elementary and middle school students. There is also a myriad of assessments that assess the application of energy ideas in different contexts, such as in thermodynamics (Wattanakasiwich, Taleab, Sharma, & Johnston, 2013) or in different science disciplines (Lee & Liu, 2009). These targeted assessments can be too specific for teachers who need a broader assessment of what their students know about energy. What is needed is an assessment that can be taken by both basic and advanced learners and that provides an overall picture of what a student knows about energy.

In the work reported here we developed an assessment tool for measuring students' growth in understanding energy from 4th to 12th grade. Using a bank of hundreds of energy items, three assessment instruments were created for measuring students' understanding of 14 energy ideas. These three instruments were pilot tested, and student responses were modeled using Rasch modeling. Results were used to verify that the three instruments performed reliability, defined a common scale, had the appropriate range of difficulty for the target grade band, and adequately tested students' growth of understanding across the 14 energy ideas at progressively higher levels of sophistication. Our results showed that after minor edits the three instruments form a valid vertical test framework for assessing students' understanding of energy.

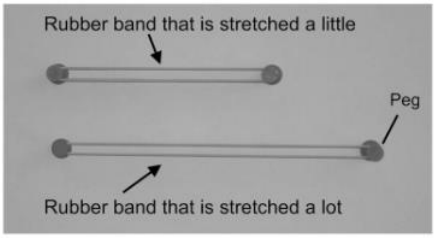
Methodology

Instrument Development. Development of the tests began with explicitly defining the construct to be tested. Four basic energy categories are typically described by researchers in this area (see, for example, Duit, 2014) including (1) energy forms and transformations, (2) energy transfer, (3) energy dissipation, and (4) energy conservation. We identified fourteen energy ideas within these categories. For example, the energy forms and transformation category included ideas about the five forms of energy (e.g. kinetic energy and gravitational potential energy) and the idea of energy transformations. The energy transfer category included six specific mechanisms of energy transfer (e.g. conduction and radiation).

Because the goal of the assessments is to be able to measure the progress of students' understanding of the energy concept, we needed to articulate a learning progression for the energy ideas that students would progress through from fourth through twelfth grade. For each energy idea, we wrote clarification statements on what student should know and articulated learning progressions with three levels of conceptual complexity. The *basic* level involves simple energy relationships or phenomenological understanding, the *intermediate* level involves using more detailed energy-concepts to explain phenomena, and the *advanced* level involves the most complex energy concepts, often requiring an atomic/molecular model to explain phenomena. The development of these learning progressions was informed by guiding documents such as *Benchmarks for Science Literacy* (AAAS, 1993), *Atlas of Science Literacy* (AAAS & NSTA, 2007), *A Framework for K-12 Science Education* (NRC, 2012), and the *Next Generation Science Standards* (NGSS Lead States, 2013). The clarification statements and learning progression was reviewed by content and education experts.

Items were created for each level of the learning progression for each energy idea. A total of 372 distractor-driven, multiple-choice items aligned to the three levels of each idea were created. Distractors were written to incorporate the current literature on energy misconceptions. For a detailed description on the item development procedure see Herrmann-Abell & DeBoer, 2014. Items were field tested in 2015 (Herrmann-Abell & DeBoer, under review). An example item is shown below:

A student has two identical rubber bands. She stretches each rubber band around two pegs so that one rubber band is stretched a little bit and the other rubber band is stretched a lot.



When the rubber bands are stretched, which rubber band has more elastic potential energy?

A. The rubber band that is stretched a little has more elastic potential energy.
B. The rubber band that is stretched a lot has more elastic potential energy.
C. The rubber bands have the same amount of elastic potential energy no matter how much they are stretched.
D. Neither rubber band has any elastic potential energy.

Items were then chosen to create three tests designed to assess the three conceptual complexity levels (*basic, intermediate, and advanced*). Each test, consisting of 35 items, was designed to be completed within an hour long class period. All tests included five linking items so that item and student characteristics could be placed on a common scale and compared across the three different instruments. Three items linked the *basic* and *intermediate* tests, three items linked the *intermediate* and *advanced* tests, and two items linked all three forms. Table 1 summarizes the number of items on each test per energy category and number of linking items.

Table 1: *Number of items by topic and level*

Test	Energy Category				Number of Linking Items
	Forms of Energy	Energy Transfer	Energy Dissipation	Conservation of Energy	
Basic	10	19	2	4	5
Intermediate	16	14	3	2	8
Advanced	15	15	2	3	5

Participants and Data Collection. The three tests were administered during the winter of 2015-2016 to 1,312 elementary, middle, and high school students throughout the United States. Elementary students (Grades 4 and 5) made up 15% of the sample, middle school students (Grades 6 through 8) 42% of the sample, and high school students (Grades 9 through 12) 43% of the sample. 53% of the students were female and 47% were male. 8% of the students indicated that English was not their primary language. All students who participated in the study were currently enrolled in a science class, but students were not necessarily being taught the target energy concepts at the time of testing. Tests were administered in paper-and-pencil format, and students bubbled in answer choices using a separate answer sheet.

In order to collect data from students with a wide range of knowledge about energy, each test was given to multiple grade bands. The *basic* and *intermediate* tests were administered to all grade bands (elementary, middle, and high school students), and the *advanced* test was administered to middle and high school students. Elementary school students were excluded from the *advanced* test because it assessed concepts, such as atoms and molecules, usually not introduced till middle school.

Rasch Analysis. Rasch analysis was conducted using the software package WINSTEPS (Linacre, 2016). In the Rasch model, the probability of a student answering an item correctly is a function of that student's knowledge and the item's difficulty. To improve the data's fit to the Rasch model, students who answered fewer than six items, and students with item Z-residual statistics higher than 4, were excluded from Rasch analysis. Answering a low number of items may indicate the student was not taking the assessment seriously, and having a high Z-residual may indicate the student was guessing. Removal of these data resulted in the final data set consisting of 1,286 students.

Linking Item Analysis. The three tests were created by assigning items based on the cognitive complexity of the targeted content and the item difficulties that had been observed during the field testing of the items that took place in the spring of 2015 (Herrmann-Abell & DeBoer, under review). The three tests were considered to be on a common scale because of the use of linking items.

To validate that the linking items performed similarly on the three different tests, we conducted a differential item functioning (DIF) analysis for each linking item. ETS guidelines were used to classify whether a linking item had *negligible*, *slight to moderate*, or *moderate to large* DIF (Zwick, 2012). Items with *slight to moderate*, and *moderate to large*, DIF were further examined by modeling each test separately and then cross plotting item difficulties (Figure 1). Items that perform similarly on the different test forms should lie, within error bars, on a line with slope approximately equal to one.

Unidimensionality Analysis. We also tested the three instruments for unidimensionality by performing a Principle Component Analysis (PCA) on the items' standardized residuals using WINSTEPS.

Results

Fit to the Rasch Model: Table 2 summarizes the fit statistics for both the items and students. All items had acceptable infit and outfit mean-square values and positive point-measure correlations, indicating a good fit to the Rasch model. The high item separation index indicated that the test accurately differentiates between levels of item difficulty. The student separation index is lower than the item separation; however this is likely due to our use of matrix sampling which meant that students took only a subset of the total set items.

Table 2: *Summary of Rasch Fit Statistics*

	Item			Student		
	Min	Max	Median	Min	Max	Median
Standard error	0.06	0.16	0.12	0.35	1.15	0.39
Infit mean-square	0.82	1.33	1.00	0.58	1.68	0.99
Outfit mean-square	0.67	1.43	1.00	0.26	2.21	0.98
Point-measure correlation	0.01	0.56	0.40	-0.5	0.9	0.33
Separation index (Reliability)	7.05 (0.98)			1.69 (0.74)		

Performance of Linking Items and Forming a Common Scale. DIF analysis indicated that two linking items may have performed differently on the different tests. One item was flagged for having *moderate to large* DIF when comparing its performance on the *basic* and *intermediate* tests; a second item was flagged for having *slight to moderate* DIF when comparing its performance on the *intermediate* and *advanced* tests. To further analyze whether this difference was meaningful we Rasch modeled each test separately and cross-plotted the linking items' difficulties. Figure 1 shows four items outside the expected linear fit (Figure 1). Both items that had been flagged for having *slight to moderate* or *moderate to large* DIF were located outside the expected fit region on the cross-plots, providing further evidence these items may not be suitable as linking items.

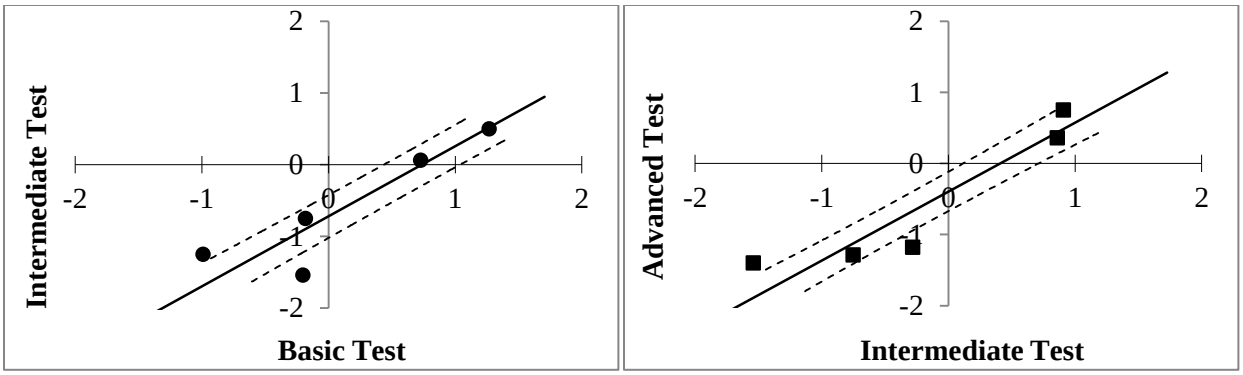


Figure 1: Cross plot of item difficulties of linking items

Based on these finding, the two linking items that were found to be outside the expected fit region were treated as non-linking items in order to test whether the linking between tests could be improved. To do this, the DIF-flagged item linking the *basic* and *intermediate* tests was removed from the *basic* test data set and used only on the intermediate test, and the DIF-flagged item linking the *intermediate* and *advanced* tests was removed from the *intermediate* test data set and used only on the advanced test. This decreased the number of linking items on the *basic* test by one, on the *intermediate* test by two, and on the *advanced* test by one.

Table 3 shows the revised Rasch Fit statistics after these changes were made and all data were modeled again. The revised data fit the Rasch model and all remaining linking items were found to have negligible DIF. Note the slight decrease in item separate from 7.05 will linking items to 6.67 when the two misfitting linking items were changed to non-linking items. These results indicated that the revised tests now formed a common scale, allowing for item difficulties and student abilities to be compared across the test levels.

Table 3: *Summary of Rasch Fit Statistics after removal of two linking items*

	Item			Student		
	Min	Max	Median	Min	Max	Median
Standard error	0.06	0.16	0.12	0.35	1.18	0.39
Infit mean-square	0.82	1.31	1.00	0.57	1.8	0.99
Outfit mean-square	0.67	1.41	1.00	0.26	2.30	0.98
Point-measure correlation	0.01	0.56	0.40	-0.5	0.9	0.33
Separation index (Reliability)	6.67 (0.98)			1.72 (0.75)		

Unidimensionality. PCA indicated that a large percentage of the raw variance was explained by the Rasch Model. Out of a raw unexplained variance of ~77%, the 1st contrast made up 1.7%, equal to an eigenvalue of 2.15 (or approximately two items). This small percentage is consistent with what we would expect from random variance for a unidimensional assessment (Smith, 1996). We also examined the five highest-load and five lowest-load items to check if there was any appreciable difference in terms of the energy ideas they assessed. If the assessment was multiple dimensional we would expect the highest-load items to assess to one construct and the

lowest-loading items to assess a different construct. We found that both sets of items contained items assessing multiple energy topics and that there was no noticeable difference in the way the construct was assessed. These results provide confidence that the assessment is unidimensional.

Comparison of Student and Item Rasch Measures. After confirming the existence of a common scale across all three assessments and testing for unidimensionality, we sought to confirm that the difficulty of each test was appropriate for each grade band it was intended to assess (Table 4). The mean item difficulty was -0.42 on the *basic* test, 0.07 on the *intermediate* test, and 0.29 on the *advanced* test, indicating a progression in difficulty from the *basic* test to the *intermediate* test to the *advanced* test. Comparing the range of difficulty levels of the three tests with each of their targeted grade bands indicates that the basic test was at the appropriate level of difficulty for elementary level students, but that the intermediate and advanced level tests were difficult for middle and high school students respectively. In fact, even the intermediate level test proved to be difficult for the high school students. This can be seen in Table 4, which shows that the mean student measure for the middle school students was lower than the mean item difficulty for the intermediate test, and the mean student measure for the high school students was lower than the mean item difficulties for both the intermediate and advanced tests.

Table 4: *Summary of Rasch measures by test form and grade band*

	Rasch Measure			
Test or Grade band	Min	Max	Mean	SD
Basic test	-1.83	1.70	-0.42	0.89
Elementary students	-2.49	2.25	-0.46	0.82
Intermediate test	-1.31	1.34	0.07	0.76
Middle school students	-2.25	4.72	-0.26	0.91
Advanced test	-1.31	1.28	0.29	0.63
High school students	-2.67	2.25	-0.10	0.89

Figure 2 shows the item map for each test. The item map for each test shows that each test contains items with a range of difficulties and the tests become increasingly difficult as one progresses from *Basic*, *Intermediate*, to *Advanced*. The item maps also highlight several differences between the tests. As seen in Table 4, the *advanced* test had the smallest variation in item difficulties. The item map shows this is due to several items between 0 and 1 logits having similar difficulties. The item maps also show that the both the least and most difficult items were contained on the *basic* test. Lastly, the item maps show there are a few regions in the map where there are gaps in difficulty.

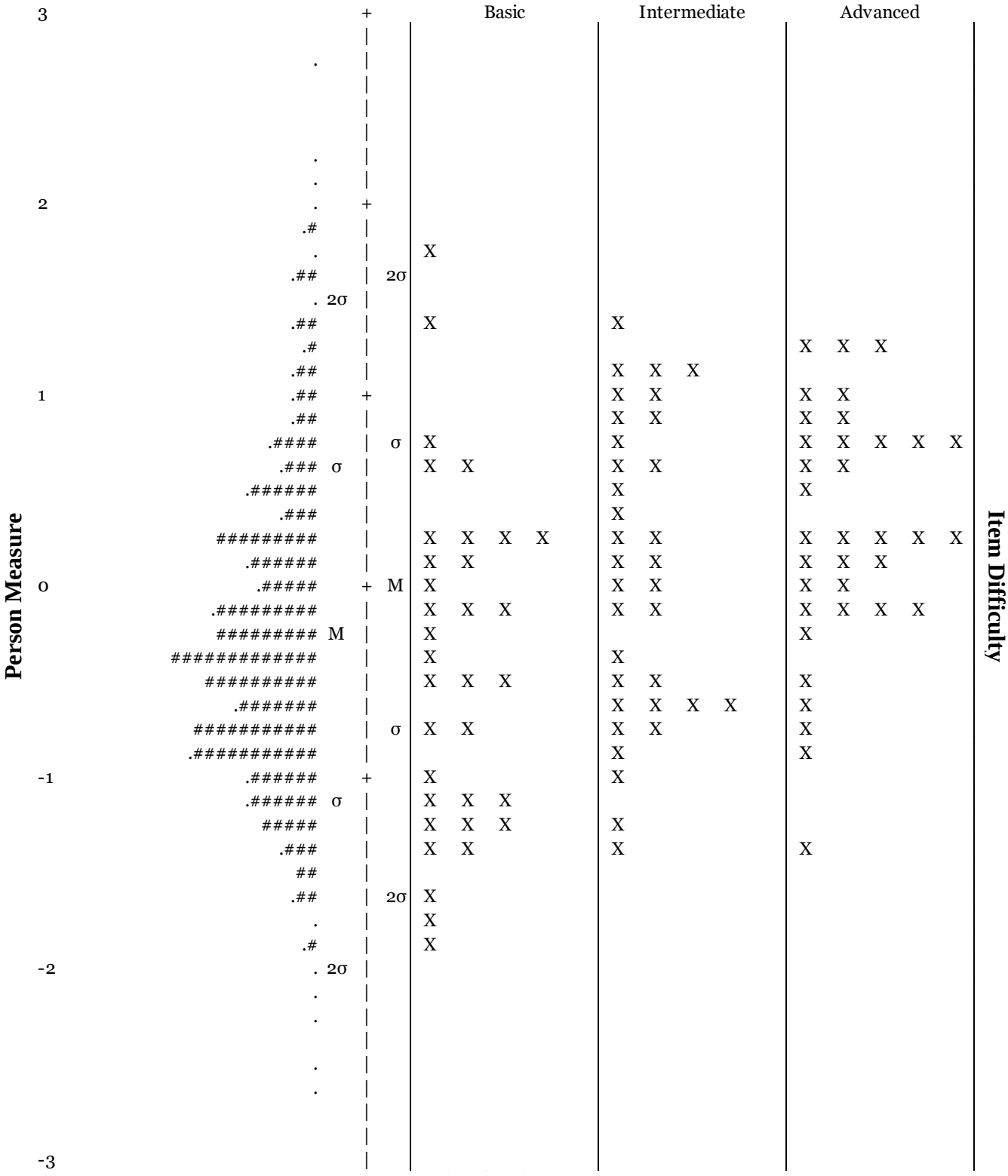


Figure 2: Item Map of *Basic*, *Intermediate*, and *Advanced* tests. X = item, # = 8 students, M = mean, σ = standard deviation, periods = 1-7 students

Discussion and Conclusions

In this work, we developed an instrument for assessing 4^{th}-12^{th} grade students' understanding of energy. The instrument consists of three tests that targeted *basic*, *intermediate*, and *advanced* energy ideas. Each instrument consisted of previously field-tested, distractor-driven, multiple-choice items that assessed students' understanding of the forms of energy, energy transfer, conservation of energy, and energy dissipation. Each instrument was administered to elementary, middle, and high school students, and the data were modelled using Rasch modeling. The modelling data was analyzed to determine whether the instrument fit the Rasch Model, was unidimensional, was appropriate for the targeted grade bands, and whether the three tests formed a common scale. Testing data was found to fit the Rasch Model; however differential item functioning identified two linking items that performed differently on the different tests. Further analysis using cross-plotting confirmed these results. This indicated that these two items should not be treated as linking items. Removing these items as linking items and modelling the data again resulted in all linking items functioning properly and the three tests forming a common scale. PCA analysis indicated the instrument measures a unidimensional construct. Lastly, the instrument consisted of items with a range of item difficulties suitable for assessing students in grades 4 through 12. Overall, the results indicate that the instrument performs reliability, forms a common scale across tests, is unidimensional, and tests students with a range of understanding of energy; however, further improvements could be made.

Our analysis indicates some areas where further improvements to the instrument could be made. Additional linking items could be added to improve the overlap of tests. There is no set rules for the number of linking items used to link tests; however, it is suggested that a test should contain 5-10 items spread across the difficulty spectrum of the test (Linacre, 2016; Raju, Edwards, & Osberg, 1983). Also, although each test contains items with a range of difficulties, the distance between the *intermediate* and *advanced* tests is small compared to the distance between the *basic* and *intermediate* tests. The item map for the *advanced* test shows some overlap in item difficulties which could be improved. More difficult items could be substituted for items with overlapping difficulties to address this. This would allow the three instruments to assess a broader range of students' understanding of energy, perhaps making the advanced test suitable for testing beyond high school. Lastly, it is worth mentioning that although the item difficulties of the *basic* test matched well with the ability measures of elementary school students, the *intermediate* and *advanced* tests are difficult relative to the ability measures of middle and high school students. One factor likely contributing to the difficulty of the *intermediate* and *advanced* tests for their targeted grade band is that these tests were designed to assess energy ideas based on recent educational standards. These standards have yet to be universally adopted, and teaching materials meeting these standards are still in development. Although these tests assess the concepts outlined in current national education standards, it would be advantageous to have middle and high school students who are being taught according to NGSS recommendations take these tests to verify that they are not too difficult for their targeted grade band after students have received appropriate instruction. We would also expect students to perform better on these tests as the standards become more widely adopted.

Implications for Educators

Given the widespread application of energy ideas, it is critical that K-12 classroom teachers and science education researchers have an effective way of measuring students' understanding of energy. The instruments developed for this study can fill this need for K-12 teachers. Results obtained from the use of these instruments can inform science instruction on the topic of energy by revealing what students know and do not know, their ability to apply that knowledge, the misconceptions they have, and how their understanding progresses from fourth through twelfth grade. Use of these instruments will also allow teachers to accurately diagnose their students' thinking, which will enable them to target instruction more effectively. Lastly, this instrument can also serve as a tool to education researchers and developers of curriculum materials. Because this instrument is carefully aligned to the key energy ideas contained in national content standards(National Research Council, 2012; NGSS Lead States, 2013) and not to any single curriculum or instructional approach, these instruments can be used to compare the effectiveness of various materials and approaches with precision and objectivity. This instrument is complementary to other energy assessments. Once a student's understanding of the general idea of energy is known, energy assessments that focus on specific energy ideas, such as the conservation of energy, can provide a more detailed picture of their understanding of the concept of energy. In addition, once a student masters the *advanced* test is may be suitable for them to take an algebra or calculus-based energy assessment, such as the Energy Concept Assessment (Ding, 2007) or the Energy and Momentum Conceptual Survey (Singh & Rosengrant, 2003).

Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A120138 to the American Association for the Advancement of Science. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

AAAS, & NSTA. (2007). *Atlas of Science Literacy*.

Ding, L. (2007). *Designing an Energy Assessment to Evaluate Student Understanding of Energy Topics*. Retrieved from https://repository.lib.ncsu.edu/handle/1840.16/4050

Duit, R. (2014). Teaching and learning the physics energy concept. In R. F. Chen, A. Eisenkraft, D. Fortus, J. Krajcik, K. Neumann, J. Nordine, & A. Scheff (Eds.). Teaching and learning of energy in K-12 education (pp. 67-85). New York: Springer.

Lee, H.-S., & Liu, O. L. (2009). Assessing learning progression of energy concepts across middle school grades: The knowledge integration perspective. *Science Education*, *94*(4), 665–688. http://doi.org/10.1002/sce.20382

Linacre, J. M. (2016). Winsteps ® Rasch measurement computer program. Beaverton, Oregon. Retrieved from Winsteps.com

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. (C. on a C. F. for N. K.-12 S. E. S. B. on S. E. D. of B. and S. S. and Education, Ed.). Washington DC: The National Academies Press.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. Washington DC: The National Academies Press.

Project 2016 (American Association for the Advancement of Science). (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

Raju, N. S., Edwards, J. E., & Osberg, D. W. (1983). The effect of anchor test size in vertical equating with the Rasch and three-parameter models. *Paper Presented at the Annual Meeting of the National Council on Measurement in Education*.

Singh, C., & Rosengrant, D. (2003). Multiple-choice test of energy and momentum concepts. *American Journal of Physics*, *71*(6), 607–617. http://doi.org/10.1119/1.1571832

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(1), 25–40. http://doi.org/10.1080/10705519609540027

Wattanakasiwich, C., Taleab, P., Sharma, M., & Johnston, I. D. (2013). Development and implementation of a conceptual survey in thermodynamics. *International Journal of Innovation in Science and Mathematics Education*.

Zwick, R. (2012). A Review of Ets Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement. *ETS Research Report Series*, *2012*(1), i–30. http://doi.org/10.1002/j.2333-8504.2012.tb02290.x