Validation of an Instrument for Measuring Students' Understanding of Interdisciplinary Science in Grades 4-8 over Multiple Semesters: A Rasch Measurement Study

Yang Yang*
Peng He
Xiufeng Liu

*Graduate School of Education
University at Buffalo, SUNY
e-mail: yang37@buffalo.edu

Paper presented at the annual meeting of the NARST – A World-wide Organization for Promoting Science Teaching and Learning through Research, April 15, 2016, Baltimore, MD

Acknowledgement: This paper is based upon work supported by the National Science Foundation under Grant No. DUE-1102998. Any opinions, findings, and conclusions or recommendations expressed in the materials are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Validation of an Instrument for Measuring Students' Understanding of Interdisciplinary Science in Grades 4-8 over Multiple Semesters: A Rasch Measurement Study

Introduction

To promote students' understanding of science, the Next Generation Science Standards (NGSS) proposes the Performance Expectations integrating three dimensions of Science and Engineering Practices (SEP), Disciplinary Core Ideas (DCI), and Crosscutting Concepts (CC) in K-12 science education (reference?). The DCIs are categorized by content areas, which include physical, life, earth and space sciences, with one more domain of engineering, technology, and applications of science. While the CC dimension contains seven concepts across subjects that help students understand the DICs deeply and develop a coherent and scientifically based view of the world (NRC, 2012). The new standards echo one of the definitions of interdisciplinarity that it is a blending of sciences where connections are made between the subjects, but they remain identifiable (Czerniak, 2007).

Furthermore, the standards remain grade specific but aligned through K-12 and emphasize learning progression along grades. For example, patterns in crosscutting concepts in K-1 life sciences require students to use patterns as evidence or in describing phenomena and in middle school life sciences the requirement is using patterns to identify cause and effect relationships (NRC, 2013). Although the advent of NGSS moved science education forward in the conceptual level, the implementation still requires great efforts from a myriad of stakeholders, such as state, district, teachers, and educators. During the process of applying the new standards, the evaluation of student learning outcomes is essential. For instance, student science achievement in terms of standardized test scores is often used as direct or indirect evidence of effectiveness of intervention, such as inquiry based instruction and teacher professional development (Guskey, 2003; Kanter & Konstantopoulos, 2010). Nevertheless, too little effort has been invested in developing more reliable, valid and engaging methods of assessment in school science (Osborne & Dillon, 2008), not to mention the assessments based on the new standards.

The evaluation of learning outcomes in science education varies according to different purposes. On one hand, large-scale standardized assessments that include international, national, and state-level, dominate from university entrance to educational policy making (Britton & Schneider, 2007). On the other hand, the small-scale assessments in classrooms take place the most often and have a stronger connection with daily teaching and learning. According to Liu (Liu, 2010a), the systematic assessments of student learning outcomes in science classroom consist of diagnostic, summative, and formative assessment based on their functions. In addition, the formats of assessments are also various. NGSS states multiple and varied assessments should be provided for students to demonstrate their competence on the expectations for a given grade level (NRC, 2014). A number of studies have showed the high reliability and validity of some alternative assessment methods in science education, such as interview, performance assessment, and concept map (Klassen, 2006; Mintzes et al. 2005; Shavelson et al. 1991; Stoddart et al. 2000; Zeidler & Sadler, 2009). However, multiple-choice (MC) questions still play a significant role in both large and small-scale assessments because of its efficiency, scoring accuracy, and economy (Roediger III & Marsh, 2005).

The content of assessments is usually in the form of integrated science for large-scale standardized exams, such as Program for International Student Assessment (PISA). While small-scale assessments mostly focus on science content knowledge, skills, and problem solving in terms of subject contexts and involving only one or few grade levels, for example, a unit test of K-7 life sciences. Therefore, both practice and study of interdisciplinary assessments related to the new standards in a school or classroom set are rare. Because MC questions remain the most common item format to assess students' learning outcomes in science (Haladyna, 2012), it is reasonable to begin evaluating student interdisciplinary science understanding by using the new standards with well designed MC items. The reliability and validity of instruments are essential in such assessments.

To improve the accuracy and effectiveness of MC questions, a large number of two-tiered instruments have been developed and validated for K-12 education to identify students' alternative conceptions of a specific science concept (Chen et al. 2013; McClary & Bretz, 2012; Sesli & Kara, 2012). Compared with simple MC questions, the two-tiered ones are able to reduce the rate of guessing and further explore student deeper understanding of science concepts. Nevertheless, those instruments mostly focus on specific science concepts or DCIs by referring to the NGSS at a targeted grade level; few of them are developed for measuring students' understanding of interdisciplinary science understanding across multiple grade levels or years. An efficient and effective way of assessing student's understanding of interdisciplinary science in multiple levels and times is called.

Students' learning outcome in integrated science is also crucial in reflecting teacher's achievement in professional development related to the new standards. As van Driel claimed (van Driel et al. 2012), measurement in student achievement is vital in the studies of effectiveness of in-service teacher professional development in implementing any curricular reform. Although the empirical research on effectiveness of interdisciplinary instruction increases in the past decade, the number is still small and the results are not consistent (Luft & Hewson, 2014). Furthermore, the measurement mostly relies on their raw scores in terms of percentage on standardized tests, such as science achievement in school report cards. Therefore, the sensitivity of measurement becomes another issue apart from reliability and validity.

The commonly used classical test theory (CTT) in developing instruments has a born limit that the instruments may not be sensitive enough to find difference between students with same scores, and changes before and after intervention (Bond & Fox, 2013). In addition, the outcome measures in CTT are dependent on the items asked and the difficulty of items is dependent on the sample used for validation (Liu, 2010b). In other words, CTT relies much on the quality of sample and items selected. Therefore, an item/sample independent method of measuring students' learning outcomes for interdisciplinary science learning is needed. Rasch modeling provides the capacity in measuring student achievement more sensitively, reliablly, and validly. One of the most important properties of Rasch measurement is the mutually independent relationship between person and item measures. It means the person ability remains the same no matter how difficult the items are and the item difficulty keeps invariant no matter what ability of students take the test. It offers the affordance in measuring difference of student ability across grades and semesters because of consistent item difficulty. The comparison of student achievement in different grades and time periods is possible by a few linking items even though they do not share exactly the same test items (Liu, 2012).

Because the CCs in NGSS are coherent throughout the K-12 expectations, it provides an opportunity to study student progression in developing understanding across grades and times. While Rasch modeling offers a tool in both validating instruments of CCs and investigating the growth of science understanding from lower to higher grade. Liu (Liu, 2007) studied students' growth from elementary to high school over an academic year in understanding the concept of matter. By using a few linking items, which the three forms (elementary, middle, and high school form) of exams shared in common, students' achievement from grade 3 to 12 and two successive semesters in the same scale were able to compare thoroughly with each other. The results showed an overall increase pattern in understanding of the concept of matter from grade 3 to 12, though fluctuating at certain grade, and no significant difference in understanding between the two semesters. The purposes of this study are: 1) to create aninterdisciplinary science instrument by using items from current reliable and valid instruments, 2) to establish evidences of sensitivity, reliability, and validity of the instrument by using Rasch modeling, 3) to improve the instrument based on the evidences, 4) to investigate student learning progression of interdisciplinary science concepts from elementary to middle school and between two successive semesters. Accordingly, research questions that guided this study are:

1. What are the empirical evidences for supporting the unidimensionality, reliability, validity, and sensitivity of the instrument for assessing students' interdisciplinary understanding of science across grade levels and times?

2. What are the evidences for suggesting further improvements of this instrument for assessing students' interdisciplinary understanding of science across grade levels and times?

3. How does students' understanding of interdisciplinary science concepts grow from elementary school to middle school and over an academic year?

Method

Instruments

The items to form the instrument for measuring interdisciplinary science understanding used in the study were selected from three sources: the Science Attitudes, Skills, & Knowledge Survey (SASKS), the Discovery's Inquiry Test (DIT), and the Ohio Achievement Assessments (OAA). The instrument aims to measure 3rd to 8th graders' understanding of crosscutting concepts (e.g. cause and effect and patterns), disciplinary core ideas (e.g. structure and function in life science), and science and engineering practices (e.g. developing and using models). To limit the length of student survey, the instrument only covers a small but significant portion of curriculum according to the new standards.

The instrument consists of 17 items. Three two-tired and one grouped MC questions are from the three forms of SASKS (each form contains a total 29 items). The two-tired MC questions focus on patterns, volume, and phase of the moon. The first item asks for the content knowledge and the follow up question requires students to explain why choosing the answer in the former one. The grouped MC question is about reading and presenting data from an experiment, in which two items share the same stem of question. The Science Attitudes, Skills, & Knowledge Survey (SASKS) was developed by Anton Lawson and the Arizona Collaborative for Excellence in the Preparation of Teachers in 2000. The instrument is designed to assess student science attitude through items of Likert-scale and measure student content knowledge and skills with multiple-choice (including two-tired questions) and open-ended questions.

Two items, one for problem solving and the other one for cause and effect, related to physical sciences are selected from DIT. The Discovery's Inquiry Test (DIT) in Science was developed in 1994 by Ohio's SSI academic leadership teams and other Ohio teachers. The test is constructed to measure student ability to analyze and interpret data and to utilize conceptual understanding of science by using NAEP 1990 and 1992 public release items. DIT includes 29 MC questions that cover four major science domains: life science (11 items), physical science (8 items), earth and space science (6 items), and nature of science (4 items). Twenty of the 29 items involved solving problems or conducting inquiry. The internal consistency of the test is high with Cronbach's Alpha of .94 (Kahle, Meece, & Scantlebury, 2000).

The rest ten items, which cover life sciences, physical sciences, earth and space sciences and scientific processes, are selected from the OAA in 2007, 2010, and 2011. The items are about patterns, scale, portions and quantity, energy and matter, structure and function, and systems and system models in perspective of the crosscutting concepts. The Ohio Achievement Assessments (OAA) is an annual statewide assessment used to measure students' knowledge of concepts taught from grade 3 to 8 in State Ohio. The purpose is to monitor and document the progress students have made in alignment with Ohio's academic standards. Science is a part of grade 5 and 8 assessments. The grade-5 science assessment from 2007 to 2011 had a high reliability of .87 (Fox, 2014).

Rasch Modeling

Rasch measurement is based on probability estimates developed by a Danish mathematician, George Rasch (Rasch, 1993). Rasch believed the probability of a person correctly responding to an item could be expressed by a mathematical equation. Reasonably, a person with high ability was more likely to respond correctly to a certain item. Likewise, the easier item was more likely to be answered corrected for person with a higher ability. He argued that the probability was governed by person ability and item difficulty simultaneously. In other words, the probability was related the difference between person ability and item difficulty. In the simplest situation, where the answer has only two possibilities, namely, correct (X=1) and incorrect (X=0), the probability, *P*, of person n correctly responding to an item i could be expressed as:

$$P(\mathrm{X}=1|\mathrm{Bn},\mathrm{Di}) = \frac{\mathrm{e}^{(\mathrm{Bn}-\mathrm{Di})}}{1+\mathrm{e}^{(\mathrm{Bn}-\mathrm{Di})}} \qquad (1)$$

Where Bn is the ability of person n and Di is the difficulty of item i. The transformation of equation (1) is shown below by taking natural logarithm for both sides.

$$\ln\left(\frac{P}{1-P}\right) = \mathrm{Bn}-\mathrm{Di} \qquad (2)$$

Where the left side of the equation is the log odds of correctly responding over incorrectly responding and the right side is the difference between person ability and item difficulty. Therefore, the likelihood of correctly responding is directly associated with the difference between ability and difficulty, and Equation (2) is the Rasch model for dichotomously scored items.

Compared with CTT, where item difficulty is estimated by percentage of correct respondents and person ability is represented by percentage of correctly answered items, Rasch measurement has two unique properties. First, person ability and item difficulty are set on a true interval scale from negative infinity to positive infinity. Second, unlike the very dependent relationship between item difficulty and

person ability in CTT, the two parameters in Rasch measurement are mutually independent as aforementioned. Therefore, Rasch measurement has two accordingly advantages in analyzing student achievement compared with raw scores. On one hand, the infinite scale in Rasch model could avoid ceiling and floor effect, which are common limits in raw score. The two effects happen when the percentage correctness reaches 100% or 0, and thus the test loses power in estimating either ability or difficulty for both ceiling and floor groups. On the other hand, the dependency between ability and difficulty hinders direct comparison between students who taking different test items, giving that students in different grades always take different tests and it is better to give different items in two successive tests. The invariance property of both item and ability in Rasch model makes it possible to create one identical scale to compare students from different tests and times by using linking items.

Nevertheless, Rasch measurement requires data to fit the model well to produce trustworthy estimates in ability and difficulty. Therefore, two assumptions and accordingly statistics should meet the requirement of a good model-data-fit. First, unidimensionality is the central assumption in single construct Rasch model in this study. It states that the items in the construct should measure the same latent trait. In practice, it is sufficient when the dominant factor exists in explaining the variance in responses (Stout, 1987). Second, the assumption of local independence states that the correlation between item responses is solely due to the examinee latent trait, thus there should not be statistically significant correlation between items in responses when person's latent trait is controlled (Liu, 2007). The local independence assumption avoids redundancy of items and the inflation of person reliability. The indices of model fit used in this study include principle component analysis of residuals, correlation of residuals, person/item reliability, INFIT/OUTFIT mean square (MNSQ), and wright map.

Confirmatory factor analysis (CFA) and analysis of variance (ANOVA)

CFA is a theory driven confirmatory technique to analyze the theoretical relationships among the observed and unobserved variables (Schreiber et al. 2006). Therefore, one function of CFA is to confirm the dimensionality of constructs according to a known theory. In this study, the data collected from the tests works as the observed variables and the theoretical dimensions are the unobserved variables. Technically, the minimum difference between the estimated and observed covariance matrices is wanted. The model of unidimensionality of understanding of interdisciplinary science through observations from the 20-item instrument is tested by using AMOS software. The ANOVA is applied in the study to investigate the significance of difference between grades and test times by comparing the means among according groups. The assumptions of independence of observations, normal distribution of data, and homogeneity of variance are explored before the implementation of ANOVA.

Sample

The number of Grade 3 students is less than 2% of the total sample and their test scores are concentrated in the low end. Thus, they are deleted in further analyses. Furthermore, students who completed blow 20% of the test are eliminated from analyses. The missing rates by using this cut point are 7%, 8%, and 8% for Fall 2013, Spring 2014, and Fall 2014 respectively. Finally, there are overall 1113 students from grade 4 to 8 in five urban schools in Northeast part of America participating and completing the two-part survey. The first part contains questionnaire on student's

background, their opinion on science, teacher instruction, and parental expectation/assistance. The second part is the 17-item instrument aforementioned. The data were collected in the three semesters with sample size of 449, 352, and 312 respectively. The overlapped students (<40%) in the first two semesters are considered as independent samples.

The distribution of student gender, race, and grade is shown in table 1. The constitution of gender is stable across three semesters (approximately half and half) with slightly more female than male students. Over one third of the participants are black in Fall 2013 and 2014, followed by white (28%) and Hispanic students (17% and 14% respectively). While in Spring 2014, there are more white students than black students who took part in the survey. Other races include multi-races, Asians, Native Indians, and Pacific Islanders. Students' raw scores in Fall 2014 have gender and race differences while in Spring 2014 there is none. The results in Fall 2013 only related to race, further analyses of gender and race are beyond the scope of this study. The number of students in 5^{th} and 8^{th} grade is relatively stable. But the number of 4^{th} graders in Fall 2014 and 7^{th} graders in Spring 2014 are too less in the matrix that might violate assumptions of statistical analyses.

Table 1 Percentages of gender, race, and grade of students in 3 semesters

	Gender		Race				Grade				
	Male	Female	White	Black	Hispanic	Others	4	5	6	7	8
F13	.47	.53	.28	.35	.17	.20	.14	.33	.16	.13	.25
Sp14	.47	.52	.38	.26	.17	.20	.13	.32	.22	.09	.24
F14	.46	.53	.28	.38	.14	.19	.02	.37	.15	.20	.27

Results

Dimensionality

The unidimensionality of the instrument was assessed by both CFA and Rasch modeling. A single construct model by connecting the items to one single latent trait was built in Amos. The model fit in CFA was evaluated by using the root mean square error of approximation (RMSEA), the relative Chi-square ($\chi 2/df$), Goodness-of-Fit statistics (GFI), and Comparative Fit Index (CFI). A good model fit is shown when RMSEA < 0.06 (Hu & Bentler, 1999).The accepted range of $\chi 2/df$ is less than 2.0 and GFI/CFI should be bigger than 0.90 (McDonald & Ho, 2002). The results of model fit in CFA are shown in table 2.

Table 2 Results of CFA for three semesters and stacked data

	$\chi 2/df$	RMSEA	GFI	CFI
Criteria	< 2.0	< .06	.90	.90
F13	1.070	.013	.970	.976
Sp14	1.270	.028	.955	.898
F14	1.319	.032	.947	.891
Stacked	1.318	.017	.984	.961

From the indexes in the table above, the data fits model well in Fall 2013 ($\chi 2/df$ = 1.070, RMSEA = 0.013, GFI = 0.970, and CFI = 0.976), while in Spring 2014 ($\chi 2/df$ = 1.270, RMSEA = 0.028, GFI = 0.955, and CFI = 0.898) and Fall 2014 ($\chi 2/df$ = 1.319, RMSEA = 0.032, GFI = 0.947, and CFI = 0.819) the comparative fit indexes are slightly lower than the criteria, which indicates possible correlation between items. However, the stacked data of three semesters illustrate good model fit

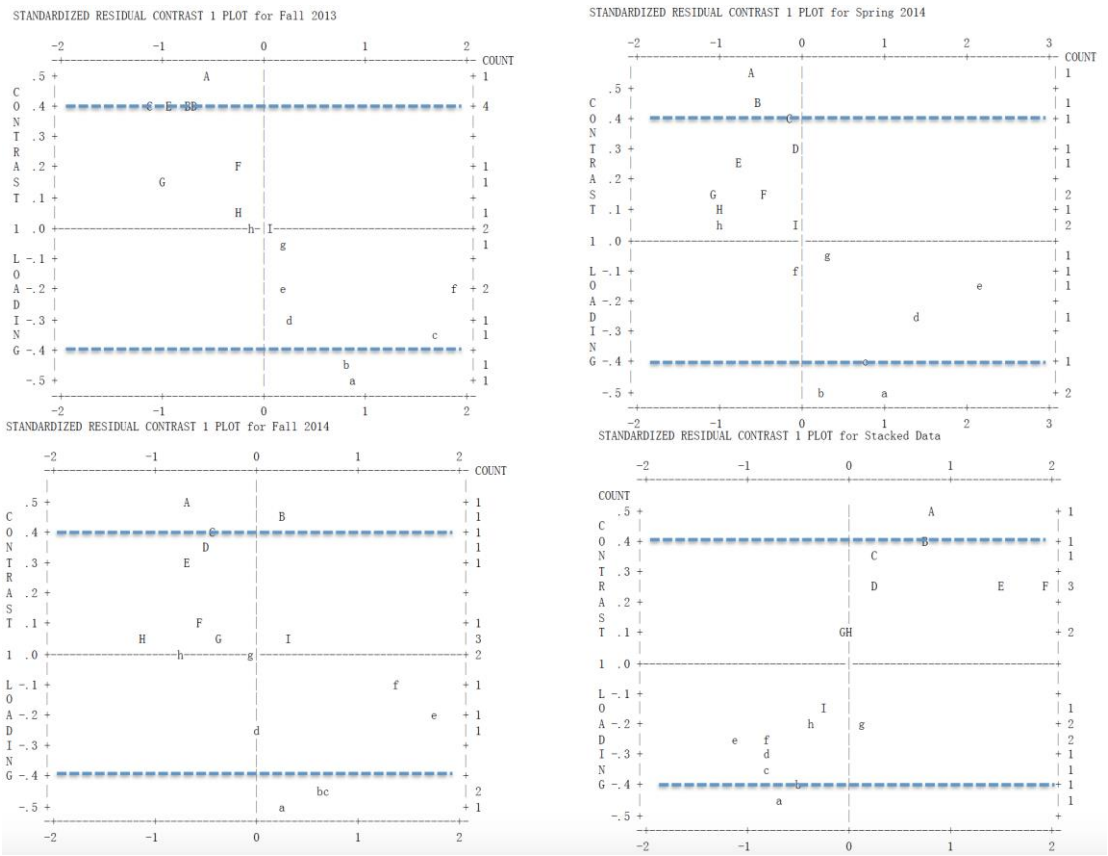
in all four aspects. Therefore, there is no significant difference between the theoretical one-dimensional construct and the observed test results.

Furthermore, the authors applied Principal Component Analysis (PCA) of residuals to identify potential additional dimensions by Rasch modeling. The total variance explained by measures ranges from 25.6% to 27.2% for the three semesters. Factor loading of items within the range of -0.4 to +0.4 indicates that those items may measure the same dimension (Liu, 2010b). The results of variance explained and items out of the acceptable range are shown in table 3 and figure 1. There are three items (E/M9, E/M11 and E/M16TT) in Fall 13, four items (E/M6, E/M9, E/M11 and E/M13) in Spring 2014, and five items (E/M4, E/M9, E/M12, EM13, E/M16TT) in Fall 2014 are out of the accepted range. To reduce the unstable variance of analyses at three different time points, students' performances in the three successive semesters are stacked into one data set. The results of stacked data show that all items are within the range of -0.4 to +0.4 except of E/M9 and E/M13.

Table 3 The results of PCA for three semesters and stacked data

	F13	Sp14	F14	Stacked
Variance explained by measures	27.2%	27.0%	25.6%	26.5%
Number of items out of range	3	4	5	2
Items out of range	EM9, EM11, EM16TT	EM9, EM13, EM6, EM11	EM4, EM13, EM9, EM12, EM16TT	EM9, EM13

Figure 1 Standardized residual contrast plots



Reliability

Rasch model estimates both item and person reliability on a scale of 0 to 1. Generally, reliabilities of 0.70 or above are considered acceptable for low stake assessment and high state tests always require a value of reliability higher than 0.90 (Nunnally et al. 1967). Item reliability depends on variance of item difficulty and sample size of respondents. Person reliability in Rasch modeling is comparable with traditional Cronbach's alpha but usually shows the lower bound of the measure (Bond & Fox, 2013). The coefficients of reliability and separation are shown in table 4. The item reliability is around 0.98, whereas the values of item separation were ranged from 5.88 to 7.50 in the three successive semesters. The high item reliability and separation indicate that the varying difficulty of items can be distinguished under the Rasch model. However, the values of person reliability are around 0.60, whereas the values of person separation are around 1.20 in the three successive semesters. Low person reliability and separation are a reflection of low power of the items in distinguishing between high and low performers (Maerten‐Rivera et al. 2015).

Table 4 Item/person reliability and separation

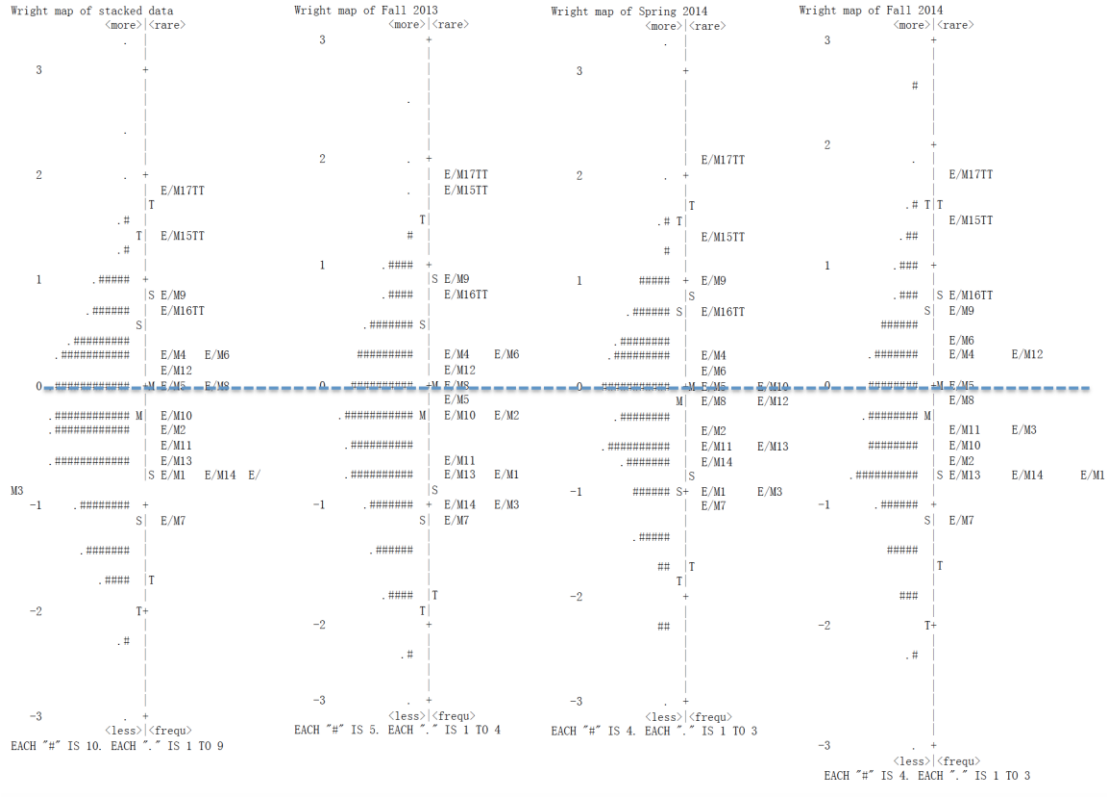
		Fall 2013	Spring 2014	Fall 2014	Stacked
Person	Reliability	0.57	0.60	0.61	0.58
	Separation	1.15	1.22	1.25	1.17
Item	Reliability	0.98	0.98	0.97	0.99
	Separation	7.50	6.76	5.88	12.15

Construct validity: Fit statistics & Wright map

The fit statistics of all items are reflected from INFIT and OUTFIT mean square. The MNSQ values (both INFIT and OUTFIT) of all items except E/M15TT and E/M17TT are within the range of 0.7 to 1.3, suggesting the items have a good model fit (Liu, 2010b). The two problematic items are with slightly higher OUTFIT mean square, which are 1.32 and 1.36 respectively, indicating less accurate predictions of the two items. The construct validity is then assessed by the person-item map in Figure 2 (from left to right: wright map for stacked data, Fall 2013, Spring 2014, and Fall 2014 respectively). It plots the relative item difficulty on the right side of the scale and personal ability estimates on the left side of the same scale. The person at the top of the map represents higher ability while the person at the bottom demonstrates low ability. Similarly, the items on the top are more difficult while the ones at the bottom are relatively easy.

The dotted line in Figure 2 is set at the average item difficulty and overall, the mean ability of students is lower than the average difficulty. Furthermore, there are major gaps at top of the maps where few students match the high difficult items and at the bottom where no items match low ability students. In addition, the authors find student abilities and item difficulties are spread widely and corresponded in range with each other consistently. Only few items show slightly variation over the three semesters, such as EM12 and EM 15, indicating a good stability of the items. Moreover, three items in the stacked map share the same difficulty, which shows a redundancy of measure by considering the total number of items.

Figure 2 Wright map for the three semesters and stacked data



Convergent validity

Evidence of convergent validity was collected by checking the association between the ability and some other parameters generated from the survey. The inter-correlations between Rasch ability and a set of well-studied parameters, such as parental expectation, and students' understanding of Nature of Science (NOS), were examined. The positive correlations between ability and parental expectation (r = 0.209, $p < 0.01$, r = 0.231, $p < 0.01$, and r = 0.271, $p < 0.01$ for three successive semesters, respectively), understanding of nature of science (r = 0.216, $p < 0.01$, r = 0.233, $p < 0.01$, and r = 0.414, $p < 0.01$) are as expected in table 5. Unexpectedly, the correlation between ability and self-efficacy is positively significant in Fall 2013 (r = 0.218, $p < 0.01$) and Fall 2014 (r = 0.156, $p < 0.05$), but it becomes non-significant in Spring 2014 (r = 0.063, $p > 0.05$). Furthermore, the Rasch calibrated ability show slightly higher coefficients than the raw scores.

Table 5 The results of inter-correlations

		Rasch Ability	Raw Score (F13)	Raw Score (Sp14)	Raw Score (F14)
Parental Expectation	F13	.209**	.196**		
	Sp14	.231**		.226**	
	F14	.271**			.263**
Students' Understanding of NOS	F13	.216**	.198**		
	Sp14	.233**		.224**	
	F14	.426**			.414**
Self-efficacy in Science	F13	.218**	.187**		
	Sp14	.063		.074	
	F14	.156*			.145*

Note: ** for $p<0.01$, * for $p<0.05$

Learning progression

The descriptive statistics of student raw scores in the three semesters are shown in table 6. Although the scores for grade 5 and 8 are relatively stable across the three semesters, the results of Fall 2014 show more heterogeneity than the previous two semesters in terms of mean, standard error, and standard deviation. In addition, students might move to a higher grade in Fall 2014, which would cause inconsistency of grade level compared with the other two semesters, and the information is not available. Therefore, the analyses of learning progression focus only on students in Fall 2013 and Spring 2014 semesters, which together represent a whole academic year. However, grade 4 and 7 maintain relatively higher errors due to their smaller sample size. To minimize the influence of sample size, stacked data of two semesters is also used to study learning progression after analyses by separated semesters. Furthermore, both Rasch abilities and raw scores are applied in progression curve to see if there is any difference in sensitivity.

Table 6 Descriptive statistics of raw scores

	N	Range	Mean (s.e.)	SD		N	Range	Mean (s.e.)	SD
Grade 4					Grade 7				
F13	60	.11-.68	.366 (.020)	.141		56	.16-.74	.496 (.021)	.139
Sp14	42	.16-.84	.468 (.024)	.142		31	.11-.74	.401 (.028)	.169
F14	6	.21-.42	.307 (.034)	.084		61	.05-.79	.455 (.023)	.178
Grade 5					Grade 8				
F13	120	.11-.79	.427 (.014)	.145		109	.11-.89	.478 (.015)	.183
Sp14	109	.11-.68	.411 (.015)	.146		81	.11-.95	.547 (.017)	.173
F14	115	.05-.84	.395 (.014)	.145		83	.05-.89	.500 (.010)	.174
Grade 6									
F13	71	.11-.74	.396 (.019)	.160					
Sp14	73	.11-.84	.441 (.018)	.160					
F14	47	.05-.74	.375 (.024)	.166					

The growth of student's raw scores across grades and times is shown in Figure 3. The mean scores fluctuate in an increasing trend from grade 4 to 8 but in Spring 2014 the 7 graders scored lower than the other students. The differences in same grade between two semesters are more obvious for 4^{th}, 7^{th}, and 8^{th} graders. The ANOVA is conducted to investigate the differences in grades and times, and the

interaction between grade and time. The Levene's test of homogeneity of mean scores shows no violation of assumption of ANOVA and thus, the results are reliable. The source table below (Table 7) shows that there is significant difference between grades (the overall means of two semesters of a certain grade) ($F = 11.91$, $p < 0.001$, $\eta = 0.06$, power = 1.00) but no significant difference between means in the two semesters. The interaction between grade and time is significant ($F = 6.41$, $p < 0.001$, $\eta = 0.004$, power = 0.99). However, the attribute of grade to the difference in raw scores is low (6%). The interaction indicates the differences in grade 4, 7, and 8 between the two semesters might be statistically significant.

Figure 3 Learning progression of Fall 2013 and Spring 2014

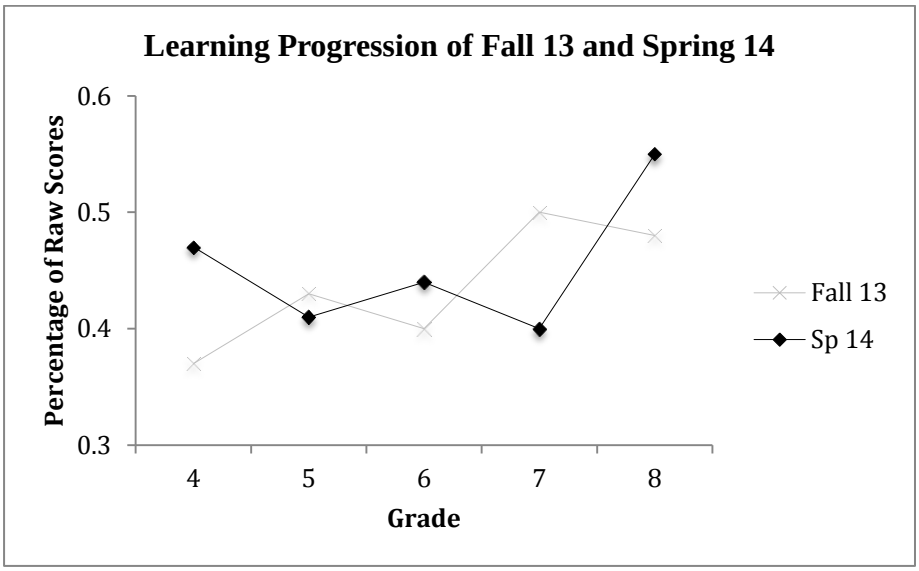


Table 7 Source table of ANOVA

Source	DF	F Values	Eta Values	P Values	Power
Grade	4	11.91	0.060	0.000	1.000
Time	1	2.87	0.004	0.091	0.394
Grade*Time	4	6.41	0.033	0.000	0.991
Testing Error	742	18.34			

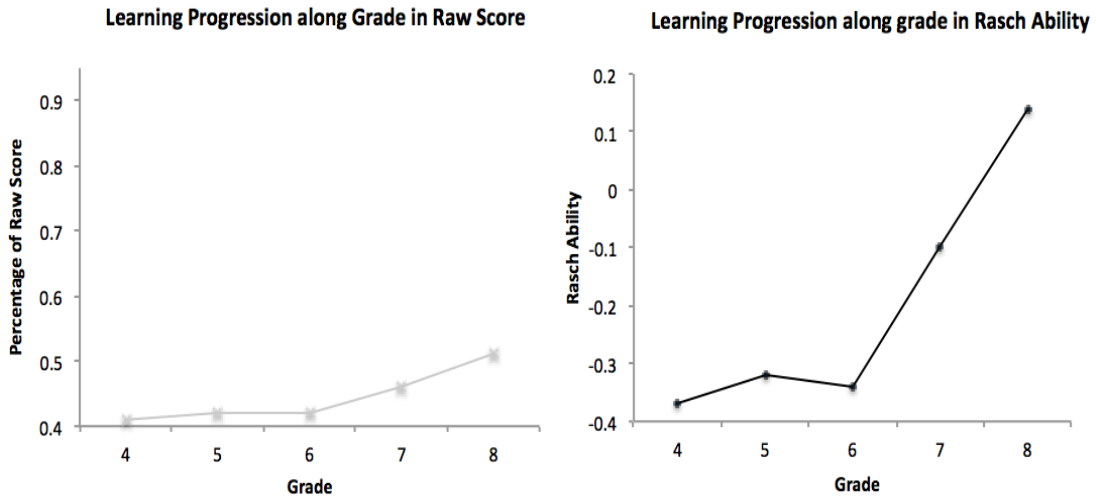
The Turkey post-hoc tests are conducted to locate the significant difference between groups. The significant difference in grade 4 raw scores between two semesters ($F = 12.95$, $p < 0.001$) becomes none significant by using Rasch ability, while the differences in grade 7 and 8 remain. Results from the stacked data show a steady increase from grade 4 to 8 and the growth rate is slightly higher in middle school than in elementary school (figure 4 left). When using Rasch calibrated ability scores, a similar trend in elementary school is found but the increase in middle school is much sharper due to different scale methods.

Table 8 Results of Turkey post-hoc tests

	Homogenous Group Means at alpha = 0.05										
	2013 Fall			2014 Spring				Combined			
Grade	N	A	B	Grade	N	C	D	Grade	N	E	F
4	60	.37		7	31	.40		4	102	.41	
6	71	.40		5	109	.41		5	229	.42	
5	120	.43	.43	6	73	.44		6	144	.42	

8	109	.48	4	42	.47	.48	7	87	.46	.46
7	56	.50	8	81		.55	8	190		.51
4					.47					
		.37								
7					.40					
		.50								
8					.55					
		.48								

Figure 4 Comparison of learning progression by using raw score and Rasch ability



Discussion

Empirical evidence of instrument quality (research question 1)

The model fit of CFA and Rasch PCA confirmed the single construct of the instrument. Thus the method of selecting items from the three sources and a variety of subject areas/concept domains is not an issue to remain unidimensionality. The relatively small CFA values in separate semester reflect the correlations among three items, which share the same stem. The two items slightly out of range in PCA are related to energy transfer and soil erosion, which indicates that they might not fit into the construct as well as the other items. Nevertheless, a distinct second trait not related to the major construct is not suggested. The question related to soil erosion (EM9) also has the highest difficulty besides of the two-tired questions. The reason of low fit of EM9 might be the unfamiliarity with the topic. Another question of energy transfer (EM13) has a proper difficulty in the instrument, however, according to previous research on students' understanding of energy, most of the students did not have a deep and conceptual understanding of the very abstract physical quantity (Harrison et al. 1999; Lee & Liu, 2010; Lewis & Linn, 2003; Nordine et al. 2011). Therefore, these two questions might not test student's understanding of disciplinary science but their guessing ability for the former and memory of energy transfer examples for the latter.

The reliability of the instrument is below the requirement of low state exam of 0.70, which indicates the items do not separate the low and high performers as expected. The reason is part from the mismatch of item difficulty and student's ability in instrument design. The distribution of raw scores, though normal, is centered on 43% with very few students got more than 80% and a number of students scored lower than the average guessing rate. The difficulty of items is over the ability of

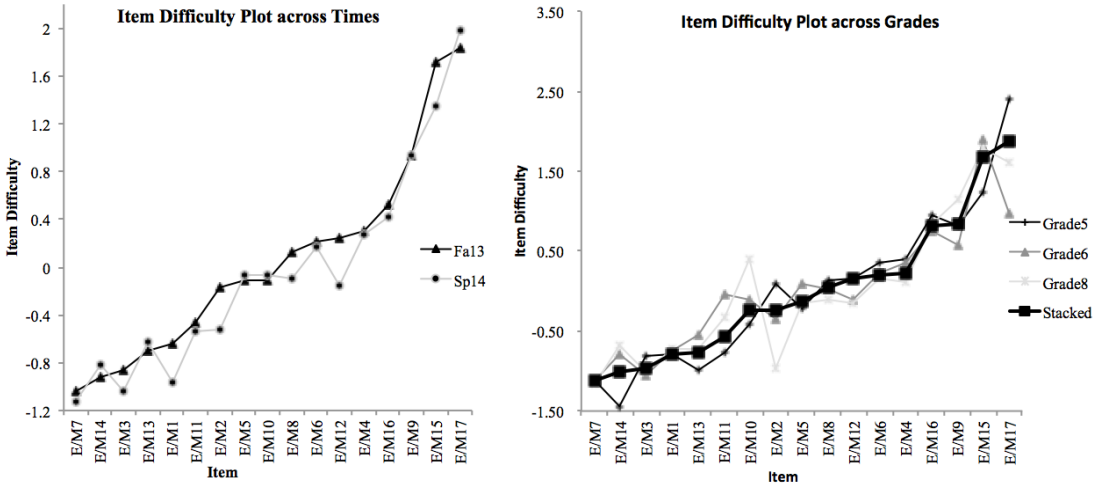
sample students. The Rasch calibrated scores illustrate the same issue, though the floor effect is eliminated. In addition, the average statewide science scores for grade 4 and 8 of the five schools in 2013 to 2014 academic year further reflect the situation. The 4th grade science average ranges from 38% to 85% and 8th grade science drops dramatically to a range of 5% to 49%. The science ability of the sample students is below the state average.

The validity of the instrument is verified through a number of methods. First, the fit statistics show that all items measure student's ability in accepted scope. In other words, the overall estimation of students' ability and item difficulty is valid and accurate. The two items slightly out of range indicate that high ability students are not more likely to respond to them correctly compared with low ability students. According to the person-item map, these two questions are the most difficulty ones, suggesting that the difficulty of them is over almost all students' ability and some correct responses might be due to guessing. Therefore, these items are not suitable for the purpose of this test. Second, the wright map provides evidence for an overall accepted instrument. The ability of students shows a great normal distribution and item difficulty spreads wide and continuous. However, There are obvious gaps at both ends. At the top of the map, almost none students have enough ability to figure out the two most difficult items. While at the bottom of the map, there are a certain number of students who are not matched with any easy items.

Finally, the correlation studies between the measure of latent trait of the instrument and a few well-studied parameters show expected results. Parental expectations in student's academic achievement, as one form of parental involvement, have been proved to be positively related to student's academic outcome with small to medium effects, including science (Davis-Kean, 2005; Hill & Tyson, 2009; Jeynes, 2007). The overall small but significant correlation between student's outcome and parental expectation is consistent with previous studies. According to Lederman (Lederman, 2007), students' understanding of NOS would enhance their subsequent learning of science subjects, but systematically empirical research on the topic was rare. The results in this study reveal positive association between students' understanding of NOS and ability in disciplinary science. Student's self-efficacy in science is positively related to their science achievement in general (Baker & White, 2003; Britner & Pajares, 2001; Meluso et al. 2012), however, the effect of self-efficacy might be mediated by other parameters, such as effort regulation (Komarraju & Nadler, 2013). The positive correlations in Fall 2013 and Fall 2014 are as expected, but the coefficient in Spring 2014 is not significant. By checking the scatter plot of data in the semester, the reason might be that a small number of students with relatively high scores move toward the low end of self-efficacy survey.

The sensitivity of Rasch calibrated scores is shown in two-fold. On one hand, Rasch ability is set on a true scale of infinity, it turns the differences exist in raw scores into a more reasonable way (Figure 4). In the case of comparing raw scores in grade 4 and 7 between Fall 2013 and Spring 2014, the raw scores are influenced severely by the small sample size. The test from ANOVA shows significant difference in raw scores in grade 4 while the difference becomes none significant by using Rasch ability. On the other hand, the assumption of Rasch modeling provides extra ways to evaluate quality of items. Because difficulty of items is assumed to be unchanged, a well-designed item should have consistent difficulty across grades and test times with non-significant noise. The variation of item difficulty through grades and times are shown in Figure 5, where most items have a stable performance in

different grades and times. A few items including EM1, EM2, EM12, EM10, EM14, EM15, and EM17 are worth more attention because of their relatively larger noise.
Figure 5 Item difficulty plot across times and grades



Improvement suggestions of the instrument (research question 2)

According to the above analyses and discussion, the overall performance of the instrument is acceptable. Thus, the practice of selecting items from a variety of sources to form an integrated science test and apply it in different grades and semesters is feasible. However, improvements are required for a higher stake assessment when targeting similar sample students. The suggestions of modification are in three ways, namely, adjustments of current items, refining of the instrument, and improvements of assessment design.

First, the items with relatively lower performance in Rasch analyses need to be revised (table 9). Because it seems students are not familiar with the topic in EM9, it is better to replace it with an item of similar difficulty in another topic. The examples of energy transfer often reflects the textbooks with which teachers are using (Kali et al. 2008), a more novel example might be use in the stem of item (EM13) to avoid simple recall. Rasch model illustrates that three items have same item difficulty, which reflects redundancy of measuring students' ability. It is also more likely to create gaps on the scale of item difficulty with redundancy by considering the total number of items. Thus, EM3 and EM14 should be either replaced by easier items or modified to reduce their difficulty. Because EM1 is part of a grouped MC question, it is better to be kept for integrity. The two most difficulty two-tiered questions are performed poorly in the current setting. One possible reason is student's reasoning ability is constrained within the single choice of follow-up question, although their reasons in mind are multi-directional and multi-leveled (Lawson, 1993). Any reasonable responses no matter the level or complexity, worth crediting fully or partially rather than a dichotomous feedback. Therefore, a written justification is more proper as the follow up question in explaining the reason of previous choice.
Table 9 Issues and improvements of current items

Item	Issues	Suggestions of adjustment
EM9	Outside the range of unidimensionality	Replace by a new item
EM13	Outside the range of unidimensionality	Use another example in stem

EM1, EM3, EM14	Redundancy of measure	Reduce difficulty of two items
EM2, EM12, EM10	Unstable across grades or times	Revise
EM15, EM17	Unstable across grades or times/ high difficulty	Change the reasoning part into written justification

Second, a few items would be added after the adjustments of current 17 items due to the gaps shown in the wright map. The small gaps above the average difficulty might be covered by the adjustments. The major gap at one standard deviation below the mean in the scale of item difficulty reflects the lack of easy items according to students' ability in the sample. A total number of 25 items by adding 6 easy and 2 difficult ones might be appreciate for the purpose of assessment. Third, the current design of assessment by using same instrument for 4^{th} to 8^{th} graders in different semesters could not fulfill the advantages of Rasch measure. The properties of true scale and invariant item difficulty provide more reasonable assessments by using a few anchor items, in which students in elementary and middle schools are able take different tests and students in successive semesters could also use separate forms of tests. Raw scores are invalid in comparing students' ability under this circumstance. Thus, the final step of improvements would be design of various forms of instrument according to times and groups of students.

Learning progression in interdisciplinary science (research question 3)

Students' learning progression in terms of understanding in interdisciplinary science is significant across grades but not in time. The overall findings are consistent with previous study of student's concept in matter (Liu, 2007). Furthermore, the results show almost no change from grade 4 to 6 while a more rapid growth from grade 6 to 8, which reflects the difference between elementary and middle schools in terms of conceptual understanding of certain topics. The findings are consistent with previous studies of learning progression of elementary and middles school students (Smith et al., 2006), where their understanding of certain scientific concepts is at distinguishable levels. In addition, the significant difference in grade 7 between two semesters might be due to high standard error. While the growth in grade 8 is trustworthy and further study is required to reveal the reasons.

Limitations and Implications

The limitations of this study are in three ways. First, the sample students in this study are constrained in five underperformed public schools. The results might not be generalized enough. Second, the quality of both survey questions, such as parental expectation, self-efficacy, and the instrument are in test stage. The reliability and validity of the findings might be undermined. Finally, the sample students are selected in convenience and not closely followed in each semester. Thus, the analyses are based on the assumption of independent sample in different semesters and the data is cross-sectional. A longitudinal data set would be more appropriate in studying students' learning progression (Johnson, 1998).

The implication of the study is two-fold. On one hand, it provides a method in designing, validating, and improving an interdisciplinary science tests. By using Rasch modeling, a more sensitive measurement across grades and times is available. The measure is essential in a myriad of studies, for example, the effectiveness of intervention in classroom and the efficient of teacher professional development. On

the other hand, Rasch calibrated scores are more reliable in comparing student ability in practice. Any consideration of student's achievement or progression by referring to their achievement in standardized tests should be precautious in the using of raw scores.

Reference

Baker, T. R., & White, S. H. (2003). The effects of GIS on students' attitudes, self-efficacy, and achievement in middle school science classrooms. *Journal of geography*, *102*(6), 243-254.

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*: Psychology Press.

Britner, S. L., & Pajares, F. (2001). Self-efficacy beliefs, motivation, race, and gender in middle school science. *Journal of Women and Minorities in Science and Engineering*, *7*(4).

Britton, E. D., & Schneider, S. A. (2007). Large-scale assessments in science education. *S. Abell, & N., Lederman, eds. Handbook of research on science education. Mahwah, NJ: Lawrence Erlbaum Associates, Inc*, 1007-1040.

Chen, Y.-L., Pan, P.-R., Sung, Y.-T., & Chang, K.-E. (2013). Correcting misconceptions on electronics: Effects of a simulation-based learning environment backed by a conceptual change model. *Journal of Educational Technology & Society*, *16*(2), 212-227.

Czerniak, C. M. (2007). Interdisciplinary science teaching. *Handbook of research on science education*, 537-559.

Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: the indirect role of parental expectations and the home environment. *Journal of family psychology*, *19*(2), 294.

Fox, A. R. (2014). *EXAMINATION OF CONSISTENCY ON THE OHIO ACHIEVEMENT ASSESSMENTS AND OHIO GRADUATION TEST.* Marshall University.

Guskey, T. R. (2003). What Makes Professional Development Effective? *The Phi Delta Kappan*, *84*(10), 748-750. doi:10.2307/20440475

Haladyna, T. M. (2012). *Developing and validating multiple-choice test items*: Routledge.

Harrison, A. G., Grayson, D. J., & Treagust, D. F. (1999). Investigating a grade 11 student's evolving conceptions of heat and temperature. *Journal of Research in Science Teaching, 36*(1), 55-87.

Hill, N. E., & Tyson, D. F. (2009). Parental involvement in middle school: a meta-analytic assessment of the strategies that promote achievement. *Developmental psychology*, *45*(3), 740.

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal, 6*(1), 1-55.

Jeynes, W. H. (2007). The relationship between parental involvement and urban secondary school student academic achievement a meta-analysis. *Urban education, 42*(1), 82-110.

Johnson, P. (1998). Progression in children's understanding of a 'basic'particle theory: A longitudinal study. *International Journal of Science Education, 20*(4), 393-412.

Kahle, J. B., Meece, J., & Scantlebury, K. (2000). Urban African-American middle school science students: Does standards-based teaching make a difference? *Journal of Research in Science Teaching, 37*(9), 1019-1041.

Kali, Y., Linn, M., & Roseman, J. E. (2008). *Designing Coherent Science Education: Implications for Curriculum, Instruction, and Policy. Technology, Education--Connections (TEC) Series*: ERIC.

Kanter, D. E., & Konstantopoulos, S. (2010). The impact of a project-based science curriculum on minority student achievement, attitudes, and careers: The effects of teacher content and pedagogical content knowledge and inquiry-based practices. *Science Education, 94*(5), 855-887.

Klassen, S. (2006). Contextual assessment in science education: Background, issues, and policy. *Science Education, 90*(5), 820-851.

Komarraju, M., & Nadler, D. (2013). Self-efficacy and academic achievement: Why do implicit beliefs, goals, and effort regulation matter? *Learning and Individual Differences, 25*, 67-72.

Lawson, A. E. (1993). Deductive reasoning, brain maturation, and science concept acquisition: Are they linked? *Journal of Research in Science Teaching, 30*(9), 1029-1051.

Lederman, N. G. (2007). Nature of science: Past, present, and future. *Handbook of research on science education*, 831-879.

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.

Lewis, E. L., & Linn, M. C. (2003). Heat energy and temperature concepts of adolescents, adults, and experts: Implications for curricular improvements. *Journal of Research in Science Teaching, 40*, S155-S175.

Liu, X. (2007). Elementary to High School Students' Growth over an Academic Year in Understanding Concepts of Matter. *Journal of Chemical Education, 84*(11), 1853.

Liu, X. (2010a). *Essentials of science classroom assessment*: Sage Publications.

Liu, X. (2010b). *Using and developing measurement instruments in science education: A Rasch modeling approach*: Iap.

Liu, X. (2012). Using Learning Progression to Organize Learning Outcomes: Implications for Assessment. *Making it tangible: learning outcomes in science education*, 225-241.

Luft, J., & Hewson, P. (2014). Research on Teacher Professional Development Programs in Science. *Handbook of research on science education, 2*, 889-909.

Maerten-Rivera, J. L., Huggins-Manley, A. C., Adamson, K., Lee, O., & Llosa, L. (2015). Development and validation of a measure of elementary teachers'

science content knowledge in two multiyear teacher professional development intervention projects. *Journal of Research in Science Teaching, 52*(3), 371-396.
McClary, L. M., & Bretz, S. L. (2012). Development and assessment of a diagnostic tool to identify organic chemistry students' alternative conceptions related to acid strength. *International Journal of Science Education, 34*(15), 2317-2341.
McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological methods, 7*(1), 64.
Meluso, A., Zheng, M., Spires, H. A., & Lester, J. (2012). Enhancing 5th graders' science content knowledge and self-efficacy through game-based learning. *Computers & Education, 59*(2), 497-504.
Mintzes, J. J., Wandersee, J. H., & Novak, J. D. (2005). *Assessing science understanding: A human constructivist view*: Academic 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.
NRC. (2012). *A Framework for K-12 Science Education:: Practices, Crosscutting Concepts, and Core Ideas*: National Academies Press.
NRC. (2013). Next Generation Science Standards: For States, By States: National Academies Press Washington, DC.
Nunnally, J. C., Bernstein, I. H., & Berge, J. M. t. (1967). *Psychometric theory* (Vol. 226): McGraw-Hill New York.
Osborne, J., & Dillon, J. (2008). *Science education in Europe: Critical reflections* (Vol. 13): London: The Nuffield Foundation.
Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*: ERIC.
Roediger III, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 1155.
Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of educational research, 99*(6), 323-338.
Sesli, E., & Kara, Y. (2012). Development and application of a two-tier multiple-choice diagnostic test for high school students' understanding of cell division and reproduction. *Journal of Biological Education, 46*(4), 214-225.
Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied measurement in education, 4*(4), 347-362.
Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). FOCUS ARTICLE: Implications of Research on Children's Learning for Standards and Assessment: A Proposed Learning Progression for Matter and the Atomic-Molecular Theory. *Measurement: Interdisciplinary Research & Perspective, 4*(1-2), 1-98.
Stoddart, T., Abrams, R., Gasper, E., & Canaday, D. (2000). Concept maps as assessment in science inquiry learning-a report of methodology. *International Journal of Science Education, 22*(12), 1221-1246.
Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589-617.
van Driel, J. H., Meirink, J., Van Veen, K., & Zwart, R. (2012). Current trends and missing links in studies on teacher professional development in science

education: a review of design features and quality of research. *Studies in science education, 48*(2), 129-160.
Zeidler, D., & Sadler, T. (2009). Scientific literacy, PISA, and socioscientific discourse: Assessment for progressive aims of science education. *Journal of Research in Science Teaching, 46*(8), 909-921.