Subject/Problem

Disciplinary-based learning is still the norm in today's high school and college classrooms, even though many science topics in secondary and higher education are highly interdisciplinary. Traditional science assessments in national and international tests rely heavily on disciplinary-based items that require separate elements of scientific information. In this situation, students are prevented from discovering and creating links between any relevant science subjects, which eventually leads to poor interdisciplinary understanding of issues in science. The philosophy of the study is not that disciplinary learning and interdisciplinary learning are mutually exclusive. Rather, it recognizes that each discipline is also important for interdisciplinary understanding. Approaching the boundary of a discipline can be a precursor to an interdisciplinary approach and students need a base of domain knowledge before integration of the knowledge. As a result, the interdisciplinary learning focus neither blurs nor erases the disciplines; rather, it elucidates the usefulness of their distinctions by clarifying their universal connections (Metz, 1995).

Interdisciplinary learning has been promoted at all levels of education but there has been insufficient empirical research in the area of assessing students' interdisciplinary understanding (Shen, Liu, & Sung, 2014). The impetus for this study is the recognition of assessment as "the 'black hole' of interdisciplinary education" in K-16 science education (Boix Mansilla, 2005, p. 18). This situation calls for more authentic assessments that emphasize interdisciplinary understanding. This study suggests thus a new starting point for developing and validating an interdisciplinary assessment in science, along with the perspective of the NGSS framework focusing on the ability to integrate content knowledge across science disciplines. Assessment of interdisciplinary understanding is expected to shed some light on an important issue for implementing interdisciplinary learning and teaching. This study aims to (a) develop a reliable and valid assessment instrument to measure high school and college students' interdisciplinary understanding across the disciplines of science for the topic of carbon cycling and (b) evaluate psychometric properties of student responses obtained from the assessment to establish the construct validity using Item Response Theory (IRT).
The guiding research questions for this study are as follows:

1) How valid and reliable is the developed assessment for measuring interdisciplinary understanding of carbon cycling?
    a. To what extent do the sources of content validity index (CVI) evidence support inferences about the items?
    b. What are the characteristics of the interdisciplinary science assessment items (unidimensionality, local independence, item fit, internal consistency etc.)?
    c. How are the items on the instrument separated by difficulty and discrimination?
    d. Do items on the instrument function differently across gender? (Does differential item functioning (DIF) occur?)
    e. What proposed theoretical model best represents the internal structure of the instrument?
2) To what extent does the development process of scoring rubrics support inferences about the interpretation of scoring and inter-rater reliability of the scoring rubric developed for constructed response items?

Methods and Results

Participants and data collection: The assessment data from 44 high school and 410 college (including four graduate) students were collected in a public high school and a public research university in Texas during the 2015-2016 school year. All test responses were collected by a web-based Qualtrics system. The 454 students showed variation in their demographic information including gender, race, and grade level. These participants ranged in grade levels from 9th grade high school students to graduate students. Of these students, 41.9% were males and 58.1% were females. The racial diversity of the participants was: White (39.0%), Asian (28.9.%), Hispanic or Latino (23.3%), African American (5.1%), Native Hawaiian or other Pacific Islander (0.4%), and other (3.3%).

The current study adopted the construct modeling framework (Wilson, 2005) to develop the items through a more systematic process and used IRT models and factor analysis to establish a more robust construct validity. Also, the current study detailed the rubric development process including the inter-rater reliability issue for constructed response items based on a systematic evaluation process.

1) Development of construct map. This study developed a construct map of interdisciplinary understanding of carbon cycling. The map consists of five levels of interdisciplinary understanding in a hierarchical fashion (i.e., No response or Irrelevant, Unidisciplinary understanding, Partially interdisciplinary understanding, Fully interdisciplinary understanding**)**

2) Item Design

a) Creation of an initial item pool: The initial item pool for the interdisciplinary assessment was created based on following three tenets. First, a pool of potential items should be aligned with performance expectation on disciplinary core ideas (DCIs) shown in the Next Generation Science Standards (NGSS Lead States, 2013). The second tenet is that assessment items should contain either unidisciplinary or interdisciplinary components of science to reveal a wide range of levels in students' interdisciplinary understanding. The third tenet involves items to address real-life problems, such as global warming and ocean acidification. Additionally, This study used 'concept maps' from content experts to reveal what they considered to be core concepts within carbon cycling. Based on the three tenets and the experts' concept maps, nine core themes for the content selection in carbon cycling were selected. They were photosynthesis, cellular respiration, decomposition, carbon reservoirs, fossil fuels, deforestation, food chain (movement of carbon and energy flow), ocean acidification, and global warming.

b) Establishing content validity: For identifying agreement among experts regarding content validity, this study used content validity index (CVI), which quantifies experts' degree of agreement (Lynn, 1986). Only 18 out of 20 items received relevance ratings of 3 or 4 by all experts according to the minimum criteria for the CVI. Since two items (item 4 and 20) did not achieve the required agreement from the experts, they were eliminated. Based on the feedback received from the experts, other items were modified by (1) changing from a constructed response (CR) item to a multiple choice (MC) item, (2) adding phrases to clarify the meaning of the questions, (3) removing extraneous information in prompts, or (4) adding two new items pertaining to the physics and earth science disciplines. This led to the second version of the instrument and it was used in the pilot test.

c) Pilot test: Through evidence of content validity of the assessment, 20 items including 12 MC items and 8 CR items were obtained. Those items were piloted with 22 middle school and 17 high school students who participated in summer programs at a public research university and an SAT class held at a Presbyterian church in Texas.

d) Reviewing and refining the item pool: As a result of the pilot test, some items were rewritten to make them clearer for students to understand, and one item asking "where most of Earth's carbon resides" was removed, as this question required only recalling facts and, in the pilot test, almost all students answered this question incorrectly. Three items were removed and two new items were added with content validity confirmation from the experts. In the final version, the number of items was 19 in total.

e) Administration procedure: During the administration process of the final items, assessment data from 44 high school students (grade 9th-12th) and 410 college/graduate students were collected during the 2015-2016 school year using a web-based Qualtrics system.

3) Outcome space (rubric development process)

Outcome space describes how individual test items are to be scored and how item scores are to be combined to yield overall test scores. Separate rubrics for each item were developed, using the following seven-step process. All seven steps comprise processes to ensure the reliability and validity of the rubrics.

a) Identifying the construct: The construct map developed in the previous stage was used as a basis to build a holistic scoring rubric. The holistic rubric categorized performance level of student's interdisciplinary understanding as "non-disciplinary", "uni-disciplinary," or "interdisciplinary".

b) Identifying levels of performance and assigning scores: The elements in the holistic scoring rubric were translated into separate analytic rubrics. The analytic rubrics were presented in a way that allowed raters to objectively determine the level of interdisciplinary understating in students' responses. The determination of the perfect score for the items depends on to what extent the items have an interdisciplinary nature according to the construct map developed in the previous stage. The rubrics consist of two different score ranges, 0-6 points for two disciplinary items and 0-8 points for six interdisciplinary items, which denotes a progression from the absence of the interdisciplinary understanding to extensive use of knowledge in different science disciplines.

c) Writing a description for each point on the rubric scale: The initial analytic rubrics were drafted based on the scientific content and analysis of student artifacts. In constructing the provisional rubrics, the highest and lowest level descriptions were decided before the middle-range ones for each item were added. Further, based on a constant comparison approach (Glaser & Strauss, 1967), we randomly selected a sample of 50 responses for each item, and carried out comparisons among the students' responses and sequenced their levels depending on their response patterns.

d) Obtaining feedback on the rubric and carrying out revisions: Five experts from Earth science, biology, physics, and chemistry participated in reviewing the preliminary rubric. The experts were asked to comment on the clarity of descriptions of ideal answers and appropriateness of the rating levels. For example, a suggestion that came from an instructor with regard to an incorrect scientific fact was to replace the sentence such as "it helps the Earth hold on to more infrared radiation from the sun, which in turn warms the climate further", to "it helps the Earth hold on to more infrared radiation reflected back to the Earth by the atmosphere, which in turn warms the climate further", indicating a possibility for improvement.

e) Rater training and scoring: two raters with biology and chemistry background received a total of four hours of training. The training included an overview of the project, explanation of the scoring guide and rubrics, and discussion of sources of bias in scoring with pre-scored student samples. During the training period, the two raters individually scored the items and compared their scores, and they then discussed any discrepancies in their scores until a consensus was reached for all scores.

f) Improving the rubric based on scoring: During five weekly meetings, the rubrics were further revised through the norming processes (Bresciani et al., 2004). If an item had less than an 80% inter-rater agreement, the rubrics were refined to provide a clearer description for the raters and then the item was rescored with the revised version of the rubrics to achieve higher reliability.

g) Inter-rater reliability: To establish the inter-rater reliability of the coded data set, the percentage agreement and the intra-class correlation coefficients (ICCs) based on a two-way mixed model were evaluated. Table 1 summarizes the percentage agreement and ICCs. The percentage agreement between raters for all eight CR items was greater than 90%, ranging from 90.09%-94.7%. The ICCs for individual items ranged from 0.980 (item 15) to 0.996 (item 8). The average ICC for the 8 CR items was 0.990 with all items demonstrating excellent inter-rater reliability. Based on the results, the rubrics are considered reliable to evaluate students' performances.

Table 1. Percentage agreement and intra-class correlation coefficients assessing inter-reliability.

CR Items	Percentage agreement (%)	ICC
2	92.51	0.991**
4	94.71	0.989**
7	91.63	0.987**
8	94.05	0.996**
12	92.51	0.992**
14	90.53	0.993**

15	90.09	0.980**
19	92.95	0.994**

Note: ** Significant at 0.001 level

4) Data analyses

a) Descriptive statistics: Table 2 shows descriptive statistics according to the demographic information.

Table 2. Demographic information and descriptive statistics (N=454)**.**

	N	Percent	Min	Max	Mean	SD
Grade level						
High school	44	9.6	4	46	11.39	9.01
College/Graduate	410	90.4	4	56	30.29	10.89
Overall	454	100	4	56	28.46	12.09
Gender						
Female	264	58.1	4	55	28.47	11.50
Male	190	41.9	4	56	28.44	12.89
Overall	454	100	4	56	28.46	12.09
Race						
White	177	39.0	4	54	29.73	11.56
American Indian or Alaska Native	0	0			NA	NA
Asian	131	28.9	4	55	31.58	11.51
Native Hawaiian or Pacific Islander	2	0.4	17	24	20.50	4.95
Black or African American	23	5.1	4	40	24.70	10.91
Hispanic	106	23.3	4	56	23.81	12.99
Other	15	3.3	9	51	25.80	11.21

b) **Unidimensionality:** Prior to the main Item response analyses, the unidimensionality assumption was evaluated for the scale, using both confirmatory factor analysis (CFA) and exploratory factor analysis (EFA). In CFA, the normed chi-square value was 1.61, indicating an excellent fit (Hooper, Coughlan, & Mullen, 2008). Other fit indices, CFI= 0.932, TLI=0.924, RMSEA= 0.037 with a 90% confidence interval between 0.028 and 0.045, and WRMR=0.925, indicate that the one-factor model fits the data well (Hu & Bentler, 1999). This result provides empirical evidence that a single latent trait sufficiently explains the item responses. In EFA, a scree plot showed additional evidence of the dominance of the first factor, indicating sufficient unidimensionality.

c**) Local independence**: Values of the LD X^2 were the threshold of 10, which is considered positive evidence of local independence (Teresi et al., 2015). All LD X^2 statistics values of the ISA were relatively small, ranging from -1.6-5.3, which indicates no evidence of local dependence.

d) **Item fit:** $S\text{-}X^2$ item-fit statistic suggested by Orlando & Thissen (2000, 2003) expresses the degree of fit or misfit in items. A statistically significant difference between observed and modeled values of $S\text{-}X^2$ statistics indicates misfit for items having *p*-values less than 0.05. Only item 5 among 19 items had a *p*-value less than 0.05, which indicates a misfit item. Item 5 is a MC question asking the student to fill in the two blanks in the excerpt provided with a common gas, with the correct answer being CO_2. The ability level (theta value) of all students is higher than item 5's difficulty level, which allows us to expect that all students should get the item right. However, 7.7% of the students answered this item incorrectly. Thus, the unexpected observation of students with a higher ability level than item 5's difficulty level getting the wrong answer generated a misfit item (Boone, Staber, & Yale, 2014).

e) **Item difficulty and discrimination:** The item difficulties ranged from -2.74~0.99 across the MC items. Item 18 is the most difficult item while item 5 is the easiest item. Items 1, 3, 5, 6, 13, 16, and 17 with negative item difficulty are easy items. Items 2, 4, 7, 9, 10, 11, 12, and 14 have relatively medium difficulty because the probability of correct response is low at the lowest ability levels. Items 8, 15, 18, and 19 represented hard items. The discrimination ranged from 0.30 (Item 14) to 1.08 (Item 5). Item 14 with low discrimination shows the probability of a correct response at low ability levels is nearly the same as it is at high ability levels. Item 5 had a high level of discrimination where the probability of a correct response changes very rapidly as ability increases.

f) **DIF:** Item 6 shows non-uniform DIF; the gender difference in item 6's difficulty changes across the ability continuum, and the discrimination parameter for this item varies across gender. Item 14 shows uniform DIF across gender, indicating that this item is systematically more difficult for female groups than for male groups with the same ability. The items with DIF may cause bias, which in turn may lead to a negative impact on the construct validity.

g) **Reliability** The reliability (internal consistencies) of the items was assessed by Cronbach's alpha. The internal consistency for all 19 items was 0.782. This shows modest reliability for the instrument, according to Nunnally & Berstein (1994).

Contributions to the Teaching and Learning of Science and to NARST members

This study argues that interdisciplinary understanding is one of the overarching components that students should develop when learning science. In order to measure this interdisciplinary understanding of students in the field of science and STEM education, this study presents a new instrument for measuring interdisciplinary understanding in science that cannot be directly assessed with traditional tests. The interdisciplinary assessment could provide a powerful impetus to curriculum and instruction, stimulating changes in curriculum policy and guiding the professional development of teachers. Furthermore, the interdisciplinary assessment focusing on science and its conceptual framework can be modeled as a precursor to be applied and even expanded in the areas of technology, engineering, and mathematics.

Selected References

Boix Mansilla, V. (2005). Assessing student work at disciplinary crossroads. *Change: The Magazine of Higher Learning, 37*(1), 14-21.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer Netherlands.

Bresciani, M. J., Zelna, C. L., & Anderson, J. A. (2004). *Assessing student learning and development*. United States: NASPA.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382-386.

Metz, K. E. (1995). Reassessment of Developmental Constraints on Children's Science Instruction. *Review of Educational Research, 65*(2), 93-127.

Orlando, M., & Thissen, D. (2000). New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50-64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*, 289-298.

Shen, J., Liu, O. L., & Sung, S. (2014). Designing interdisciplinary assessments in sciences for college students: An example on osmosis. *International Journal of Science Education, 36*(11), 1773-1793. doi:10.1080/09500693.2013.879224

Teresi, J. A., Ocepek-Welikson, K., Ramirez, M., Kleinman, M., Ornstein, K., & Siu, A. (2015). Evaluation of measurement equivalence of the Family Satisfaction with the End-of-Life Care in an ethnically diverse cohort: Tests of differential item functioning. *Palliative Medicine, 29*(1), 83-96. doi:10.1177/0269216314545802

Wilson, M. (2005). *Constructing Measures : An Item Response Modeling Approach*. Mahwah, N.J.: Lawrence Erlbaum Associates.