Development and Validation of a “Scale” to Place Students along a Learning Progression

Namsoo Shin & Shawn Y. Stevens
University of Michigan, 610 E. University Ave., Ann Arbor, MI
Email: namsoo@umich.edu, sstevens@umich.edu

Abstract: The learning progression (LP) is a construct recently adopted by the science education community as a potential way to create a coherent science curriculum to support student understanding of a few core ideas. However, the field does not yet have sufficient evidence-based LPs for educators to use. We describe efforts that are part of a project that aims to develop and empirically test a hypothetical LP related to the transformation of matter for grade 6–12 students. In particular, we present the design research methodology and progress towards developing and validating an instrument developed based on a LP. The resulting instrument can locate where students fall along a hypothetical LP and characterize how a variety of factors (e.g., school, context, gender) affect students’ performance. These results will inform the development of curriculum materials and instruction that better support student understanding of ideas important for science literacy.

Recently, science educators have been working to change the focus of the curriculum to the development of an integrated understanding of a relatively small number of core ideas instead of a shallow coverage of a broad range of isolated concepts (Duschl, Schweingruber, & Shouse, 2007). To further this effort, the science education community has begun to employ the idea of learning progressions (LPs) to put emphasis on how understanding of these core ideas develops over an extended period of time. (National Research Council, 2010; Smith et. al., 2006). LPs are research-based descriptions of how students build their knowledge, and gain more expertise within and across disciplines over a broad span of time, that provide a means of organizing and aligning the science content, instruction and assessment. As of yet, the field does not have sufficient evidence-based LPs for educators for use in educational settings.

Our research group is working to develop and empirically test a LP for grade 6–12 students that emphasizes characterizing and measuring the development of connections between ideas within and across topics that are required to develop understanding of transformations of matter. The main research goal is to develop and validate an assessment to act as a ruler to place students on the hypothetical LP that can discriminate students’ performance by a variety of factors (e.g., school, curriculum, gender). We use a principled and systematic design process to guide the development and validation of assessments associated with the LP. The developed instruments are intended to monitor how grade 6-8 students may develop conceptual understanding of and apply concepts related to the structure, properties and behavior of matter to explain transformation phenomena. This research will inform the revision of the hypothetical LP as well as the associated assessment items and enable the empirical testing of a portion of the LP.

Theoretical Framework

Learning Progressions

A LP addresses a range of content within a discipline defined by a lower and upper anchor (Duschl, et al., 2007). The lower anchor provides a description of the knowledge and reasoning that should be held by students prior to beginning to develop understanding of concepts contained in the LP. The knowledge and skills students are expected to develop by the end of the progression, the upper anchor, are drawn from what educational research has defined to be developmentally feasible, and are also related to goals of science literacy and societal needs and expectations (Mohan, Chen, & Anderson, 2009). Each step, or level, along the progression should be logical, comprehensible and measurable. Many factors determine the path that students may follow, including the context, instruction, curriculum materials, and students’ prior knowledge and experiences. Therefore, we expect there could be multiple paths between levels defined on a LP. An instrument designed to measure progress along a LP must be flexible enough to monitor progress along multiple paths.

Explaining many important phenomena requires incorporating ideas from many topic areas (e.g., structure of matter, energy, conservation). We hypothesize that focusing on the growth of sets of ideas instead of individual topic areas over the course of the LP, will better support the development of integrated knowledge structures (Stevens, Delgado, & Krajcik, 2010), which in turn will result in learners who can better select and apply the appropriate ideas more consistently to explain a broader range of phenomena and to solve new problems. Using this model, levels of the LP do not progress in sophistication through the addition of a new type of idea (e.g., adding conservation or energy to understanding of the structure of matter) as every level
incorporates multiple topic areas. Thus, assessments must measure how students use and connect the ideas contained on each level of the LP.

**Construct-Centered Design Process (CCD)**

The process and methodology of the development of LP should be iterative and process-oriented for investigating student learning over a long period time. A reliable and valid instrument is necessary to track how students develop competencies over time as we improve teaching and learning strategies, and develop new materials. The complex process of developing and empirically testing a LP requires the collection, analysis and organization of an extensive amount of data, which is consistent with a design research approach (Collins, Joseph, & Bielaczyc, 2004). Such research involves the development of products (e.g., learning progression, instructional materials, technology tools) and explores the relevance of the products on learning rather than simply examining isolated variables within laboratory contexts (Brown & Collins, 1992; Barab & Squire, 2004).

The Construct-Centered Design (CCD) process (Shin, Stevens, & Krajcik, 2010) provides an approach flexible enough to guide the development of a diverse set of research products, including assessments, instructional materials and the LP itself using a design research approach. The first step of the CCD process involves defining the construct(s). A construct includes the ideas or concepts that we wish to learn about and measure (Wilson, 2005). This process involves explicitly specifying all of the content that is necessary to develop understanding of the aspects of the construct to be addressed by the LP. This also includes identifying potential student difficulties and alternative ideas related to the specified content, which is useful for creating distracters during item development. The next step involves developing a set of claims and defining the evidence that support the claims based on the relevant content for each of the constructs (Mislevy, et al., 2003; Mislevy & Riconscente, 2005). The claims describe what students should be able to do with the knowledge at a particular level in the progression. The evidence describes the ideas, skills and connections within and between constructs that students need in order to demonstrate that they have a particular level of understanding. The claims and evidence guide the development of the LP and the associated assessment tasks.

**Methodology**

Design research is by nature an iterative process. As such, there are several phases in the development an empirical testing of a LP. The first phase involves developing a hypothetical LP based on empirical research and the logic of the discipline. Cross-sectional data is collected to support and fill in any gaps in the research literature. Phase II involves the developing an instrument to measure student progress along the LP. This phase also involves collecting cross-sectional data to validate the instrument. The final phase involves empirically testing the LP through collection of longitudinal data to measure how student learning progresses. Here we discuss efforts related to Phase II.

**Developing the Instrument**

Following the model of learning and the CCD process, we developed assessment items to create a scale that will enable us to locate where students’ understanding falls along the LP. We believe that students at all levels can integrate ideas from multiple topic areas when explaining phenomena. It’s just the complexity and scientific accuracy of the models that they use in their explanations that will increase in sophistication as they move along the hypothetical LP for the transformation of matter. We tried to design items that require students to incorporate ideas from multiple topic areas regardless of the LP level with which it is associated. Middle and high school, and college students (grades 6-16; N=792) from four secondary schools and three universities in Southeast Michigan pilot tested 112 items to create a bank of items that concentrates on Levels 1-3 of the LP since that is the portion we are empirically testing (grades 6–8). More detail on this part of the process will be published elsewhere (Stevens & Shin, 2012). For the field-test data collection, we used items selected from or revised from the pilot testing. For example, we found that open-ended items generally did not provide much information. Figure 1a provides an example of an open-ended item that we modified. We found that by asking students to critique each model they did not pick encouraged them to incorporate more ideas than the original item asking them to explain why they chose their answer. On other occasions, we modified towards more closed items. For instance, instead of an open-ended item, we generated a series of true/false questions that address all of the topic areas we wished to measure so that we could obtain some information on whether students did not know how to use the ideas or just did not use them in the open-ended format. This type of item also allows us to measure multiple levels on the LP within the same item (see Figure 1b). We found multiple true/false items to be useful for other reasons as well. Higher performing students found them more difficult than typical multiple-choice items because they had to read and evaluate each statement and gained no advantage from the test-taking skills this population generally possesses. Lower performing students liked this type of item because it is not “all or nothing”. If they do not know one part, they still have a chance to answer other parts correctly so this type of item is less stressful for them. In the pilot, we gave the choice of ‘not sure’ for this type of item, but did
not include that choice in the field test. We found that the inclusion of ‘not sure’ provided cleaner data by eliminating much of the guessing issues that often accompany true/false items.

We created four test forms (1A, 2A, 3A, 4A), each containing 14 items for the 6th and 7th grade students, and four test forms (1B, 2B, 3B, 4B), each containing 15 items, for the 8th grade students. The A- and B-version of each form differ by three items to adjust the overall difficulty of the tests. Each test form contained three pairs of linking items for a total of six linking items. Linking items were chosen to cover a range of topic areas along Levels 1-3 on the LP. The A- and B-versions of each test form contained the same linking items.

Participants
We collected cross-sectional data from 37 teachers from nine schools in four different states representing a varied demographic profile including schools in urban (N = 3), suburban (N = 5) and rural (N = 1) settings. Approximately 3800 students in 6th, 7th, and 8th grades from public (N = 7), private (N = 1) and charter (N = 1) schools participated in the field test in Fall, 2010.

Data Analysis
We used 52 multiple-choice items and 10 open-ended or short-answer items in this data analysis. The total scores differed between forms because the multi-part and open-ended items had different total scores. We eliminated the students who responded to none of the items in their test. There were a total of 3405 valid subjects after this data cleaning and recapture. Data analysis was performed using ConQuest software (Wu, Adams, Wilson, and Haldane, 2007). We used Item Response Theory (IRT) to ensure the revised items were effective at measuring the construct and to create a scale to measure student progress. Differential Item Functioning analysis (DIF) and Unidimensional Latent Regression (LR) were used to characterize the different factors that could affect measurement of student performance.

Results and Discussion
Creating a scale to measure student progress
The weighted means square (MNSQ) of fit (> 1), T-value (> 3), reliability, and discrimination (< 0.3) outputs from IRT analysis were used for judging the quality of each item. In the IRT output, the fit indices indicate the extent to which the item fits the item response model. Typically, when the fit index is much greater than 1, the item is generally less discriminating than the model predicts and is therefore a potential candidate for deletion. In addition, when items exhibit a large difference between the unweighted and weighted fit value, it indicates that they lack generalizability and may have to be revised or deleted. When items did not fit the model well, we evaluated the items themselves and the scoring approach to decide whether to revise the scoring or remove the item from analysis and future data collection. We also removed items for which ≥ 90% of students answered correctly or incorrectly to account for ceiling and floor effects. After deleting the problematic items, we ran the IRT analysis again to ensure that the rest of the items fit the model well. A Wright map of the remaining items indicated that the items cover a full range of difficulty level well (-3 to 2 logits) to provide a measure of progress in 6th – 8th graders’ understanding of the transformation of matter.

Characterization and validation of the assessment
In order to evaluate the sensitivity of our instrument to various factors, we applied DIF and LR analyses. First, we wanted to ensure that our instrument could distinguish between schools. Due to the range of characteristics and contexts of the schools, we expected a large range in student performances. Using individual IRT scores, we calculated the average ability scores for students at each school (see Table 1). The higher logit values represent...
higher ability level. We see a wide range of difference in ability levels, which suggests that we can reliably discriminate between schools.

LR performed pair-wise between groups of suburban, urban and rural schools indicated that students at our urban and rural schools performed similarly (urban, 0.09 logit [SE=0.06] lower than the rural school). The students at the suburban schools on average performed over 0.3 logit higher than either the urban or rural schools. This result is consistent with our prediction that the schools with higher social economic status than the rural and the urban schools would score higher than other schools.

DIF analysis indicated that male students exhibited slightly higher, but statistically significant average ability levels than the female students (0.15 logit, SE=0.03). This difference is generally observed for middle school students (Marx, et al., 2004). We performed DIF analysis to examine differences in performance by gender at each of our schools (see Table 1). Several schools showed no significant difference in ability levels of male and female students. Interestingly, at R1 the difference was reversed with females exhibiting a significantly higher ability than male students. We will perform a detailed analysis of the curriculum materials to try to identify what might be responsible for this trend. The degree of variation between male and female performance at different schools suggests that our assessment does not have an inherent gender bias. These results provided evidence validity of the LP assessment in terms of the gender variable.

We performed LR analysis between the grades over all schools to determine whether we could measure progress along the LP using this assessment. We observed that overall, students in 8th grade showed an increase in average ability level of approximately 0.55 logit relative to the 6th graders. The increase in ability level was not uniform (0.34 logit, SE=0.02 between 6th and 7th graders; and 0.21 logit, SE=0.03 between 7th and 8th graders). To further characterize differences between schools, we compared a subset of schools. U1 and S1 both have a similar resources and science curriculum as it relates to the LP, but a very different student population. S1 and S2 have similar student populations, but differ in curriculum and resources. We observed an increase in mean ability levels at all three schools (see Table 2). As with all schools combined, the amount of progress differed from 6th to 7th grade as compared to 7th to 8th grade for the individual schools. The degree of progress at each point in the curriculum and in total differs. In addition, the total and step-wise progress differed between schools. Although the absolute ability scores appear to be correlated to SES, other factors appear to affect the degree of progress students make. We expected a smaller increase in ability level for S2 students because content associated with our LP is introduced primarily in grade 8 so at this point in the year even 8th grade students had little instruction on relevant content. In contrast, students at U1 and S1 both receive instruction on content related to the LP throughout grades 6–8. Thus, we predicted a larger increase in mean ability levels for these two schools relative to S2. A detailed analysis of the curriculum and observation of instructional practices will help us delineate the other factors affecting the students’ progress. Overall, the scores increase with grade, although there are variances among schools in the amount of learning progress and when it occurs, which is expected due to differences in context and curricula. These results suggest that our assessment can discriminate students’ performances across the grade levels.

Table 2: Comparison of increase in mean ability levels.

<table>
<thead>
<tr>
<th>School</th>
<th>6th to 7th grade (logit)</th>
<th>SE</th>
<th>7th to 8th grade (logit)</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1</td>
<td>0.55</td>
<td>0.10</td>
<td>0.35</td>
<td>0.11</td>
</tr>
<tr>
<td>S1</td>
<td>0.38</td>
<td>0.10</td>
<td>0.72</td>
<td>0.11</td>
</tr>
<tr>
<td>S2</td>
<td>0.34</td>
<td>0.06</td>
<td>0.09</td>
<td>0.06</td>
</tr>
</tbody>
</table>

*MA = mean ability logit; S = suburban; U = urban; R = rural; SE = standard error of measurement

Conclusion and Implications

This paper illustrated how we created and validated an instrument that allows us to create a detailed characterization of student performance. We will use these items examine the effect that various factors (e.g., gender, school context, curriculum materials) that can affect student learning related to explaining transformations of matter. This is a necessary step for empirically testing a LP.

Many factors (e.g., context, instruction, curriculum materials, and students’ prior knowledge and experiences) affect how and when students can learn new ideas and connect them to other related ideas. Learning research suggests that learners may follow different pathways as they work to make sense of ideas and
develop integrated understanding. Thus, multiple instructional strategies and curriculum materials may be successful in helping students to develop more sophisticated knowledge and skills, depending on their previous academic and life experiences. We will use this validated scale developed to locate students along the LP and relating student performance to determine how these factors may affect student performance. In addition, we are analyzing the curriculum materials and classroom observation data to more fully characterize students’ instructional experiences to help us connect various factors with students’ measured progress. These results will inform the development of curriculum materials and instruction that better support the development of student understanding of ideas related to the structure, properties and behavior of matter.

We are in the process of collecting longitudinal data to track student learning along the LP across three years using the validated items (Phase III). In addition, we are running confirmatory factor analysis and multidimensional IRT to measure the validity and reliability of the LP items and to determine if the items clustered as defined in the LP.

References


Acknowledgments

This work was funded by the National Science Foundation (grant numbers 0822038 and 0426328). Any opinions expressed in this work are those of the authors and do not necessarily represent those of the funding agency. We also wish to thank the participating schools, teachers and students, and our research team.

© ISLS 400