Understanding the Relationships Within and Between Constructs of a Learning Progression: Combining Multidimensional Item Response Modeling and Latent Class Analysis

Jinnie Choi and Ravit Golan Duncan
Rutgers University, 10 Seminary Place, New Brunswick, NJ 08901
Email: jinnie.choi@rutgers.edu, ravit.duncan@gse.rutgers.edu

Abstract: Learning progressions are hypothetical models of student learning in a domain over extended periods of time. In many cases these progressions describe multiple 'big ideas' or constructs. Relationships between these constructs, i.e., how development along one might affect the other, are difficult to ascertain. Such relationships can be described from the perspectives of either item characteristics or student abilities. Existing methods of analyses focus predominantly on the 'item-side' of the equation and much less research addresses construct relationships from the 'student-side'. In this study, we supplemented a Multidimensional Item Response Modeling approach with a Latent Class Analysis to more fully explore both within and between-construct relationships. We analyzed student written responses (n=317) to 31 ordered-multiple-choice items targeted at five constructs in a genetics learning progression. We present our finding with the goal of comparing and contrasting the types of inferences that can be made with both measurement approaches.

Introduction

Learning progressions (LPs) embody a developmental approach to learning by describing productive paths that students might take as they develop progressively more sophisticated ways of reasoning in a science domain over extended periods of time (Alonzo & Gotwals, 2012; Duncan & Hmelo-Silver, 2009a; NRC, 2007). Some LPs map out progress along only one core idea (e.g. Shavelson et al., 2005; Rivet & Kastens, 2012), this is termed a construct map (Wilson, 2005). Alternatively a LP can map out progress along several constructs simultaneously, thus the LP includes multiple construct maps (e.g. Duncan, Rogat & Yarden, 2009b; Jordan & Duncan, 2009; Plummer & Krajcik, 2010). A basic assumption of LPs is that within a construct map students' progress from less sophisticated levels to more sophisticated levels. However, the progress is rarely neat and linear, and diagnosing the level at which a student is reasoning can be challenging. For example, several researchers (Gotwals & Songer, 2010; Steedle & Shavelson, 2009) have pointed to the problem of a 'messy middle', in which students at the middle levels are relatively inconsistent in their reasoning on items of the same relative difficulty. This suggests the likelihood of multiple non-linear paths that students may take to reach the upper level of a LP (Steedle & Shavelson, 2009).

The relationship in progress along multiple construct maps is even more complex and can take many forms. Wilson (2009) offered several representations of how progress along multiple constructs may occur: (a) progress rates along multiple constructs may be very similar such that students progress from one level to the next along multiple constructs at the same time, i.e. the construct maps are aligned; (b) progress along one construct depends on first attaining some level of understanding along another different construct, i.e. the construct maps are staggered; or (c) two or more construct maps may 'feed' into another more sophisticated construct map, i.e. a combination of aligned and staggered maps.

Many researchers (e.g. Brown, Nagashima, Fu, Timms, & Wilson, 2010; Hadenfeldt, Neumann, & Liu, 2013; Anderson, Gotswals & Songer, 2010) have analyzed the relationships among the levels of performances within and between constructs using the multidimensional item response modeling (MIRM) approach (Adams, Wilson, & Wang, 1997; Wilson, 2013). This approach juxtaposes student abilities and individual item difficulties on the same logit scale. In the MIRM approach, inference about the validity of the proposed levels in a construct map depends mostly on how items behave given student abilities. On the 'item-side' one can calculate, for each level, the threshold point for which students have a 50% probability of achieving that level of understanding or higher (termed Thurstonian thresholds) (Wu & Adams, 2007). These thresholds are useful for inferring relative difficulties of moving from one level to the next within a construct map (Wilson & Draney, 2002). Thurstonian thresholds can also be used to infer about relative difficulties of specific levels across constructs. For example, the level one thresholds for items measuring construct X may be similar to, lower, or higher than the level one thresholds for items measuring construct Y.

On the other hand, inferences about the levels of LPs, in particular within constructs, are less informed by the 'student-side' results. The MIRMs provide student ability estimates that are normally distributed. For a five-construct test, each student gets five estimated abilities on each of the constructs and the distribution of these ability estimates for the five constructs may have different means and variances. Thus, within and between construct comparisons can be made, but only in a general distribution sense. That is, whether students can be classified into the levels of LPs is a question that MIRM approach does not directly answer. The abilities of the students are modeled and estimated to be on a continuous scale, it therefore becomes problematic to later classify students onto discrete levels of performances on LPs. This is because MIRM is based on assumption that the students are in a homogeneous group that shares a particular performance pattern on sets of assessment items; the approach assumes that all students at a particular level reason in the same, consistent, manner. Consequently, it becomes difficult to identify and understand the characteristics of the 'messy middle' classes of students. We explore whether the relationships among the levels of LP within and between constructs can be more fully explained when we supplement and bolster the MIRM approach with the missing component: providing student-side information that matches the discrete nature of the LP levels, and enables within and between construct comparisons of level dynamics in LP.

We use latent class analysis (LCA; Lazarsfeld & Henry, 1968) to provide the student-side information for our analysis of the genetic learning progression (Duncan et al., 2009b). LCA examines if the cases (e.g., students) can be placed into multiple latent groups or classes based on their response patterns. Application of LCA is not new in LP research; Steedle and Shavelson (2009) employed LCA to evaluate whether there are groups of students who perform as expected by an LP for force and motion (Alonzo & Steedle, 2009). Their results suggested that students at the lower and upper levels of the progression reasoned relatively systematically across items, however students at the middle levels often did not reason consistently and were difficult to diagnose as reasoning at a particular level. Similarly to Steedle and Shavelson we use LCA to examine whether the assumptions of our LP match the patterns of the identified classes from data. Moreover, we are particularly interested in the dual use of LCA and MIRM to provide a more complete student and item-side perspectives on the expected performance within and across the five constructs of the genetics LP. Our research questions are thus: (a) what inferences can one draw from the MIRM analysis about the relationships between levels within a construct and between levels across constructs? (b) What inferences can one draw from the LCA analysis about the relationships between levels within a construct and between levels across constructs? (c) In what ways are findings from these approaches congruent, conflicted, or enhanced by each other?

Genetics Learning Progression and Assessment Design

The genetics learning progression is organized around two core questions in the domain: (a) how do genes influence how we, and other organisms, look and function? And (b) why do we vary in how we, and other organisms, look and function? There are eight big ideas associated with these questions. In our current work we are focusing on five of them: (1) Construct A: all living things have genetic information that is organized hierarchically; (2) Construct B: the genetic information specifies proteins structure; (3) Construct C: Proteins have a central role in the biological function of living things and are the mechanism that connects genes and traits; (4) Construct E: Organisms reproduce by transferring their genetic information to the next generation; and (5) Construct F: There are patterns of correlation between genes and traits, and there are certain probabilities with which these patterns occur. Each construct is mapped out across four levels of growing sophistication. Progress along the progression entails developing more sophisticated understandings of these constructs as well how they relate to each other. A detailed description of the progression can be found in Duncan et al. (2009b).

The genetics LP, as originally described, did not provide any conjectures about how development along one or more constructs might affect development along others, as the research base was insufficient to inform such assertions. In recent work we discussed some tentative dependencies between two of the constructs (B and C) and showed that understandings of these constructs develops mostly independently and in parallel (Shea & Duncan, 2013). In this study we attempt to explore such relationships from multiple perspectives, using a larger sample, and with more powerful measurement models. Towards this end we developed a written assessment comprised of 31 ordered-multiple-choice (OMC) items corresponding to the five constructs and their four levels of understanding. In OMC items different response options are linked to levels of conceptual understanding (Briggs, Alonzo, Schwab & Wilson, 2006; Briggs & Alonzo, 2012); items are scored using partial credit models and thus provide more information about students' level of reasoning than traditional multiple-choice items.

Methods

Data and Instrument

The 31 OMC written assessment was administered, over a two-week period, by six participating teachers in 17 biology classrooms (n=317) at a suburban high school in eastern United States. The school consisted of 47% African American, 22% White, 19% Hispanic, and 11% Asian students; 34% of the students were eligible for free or reduced lunch. Among the 17 classrooms, 7 classrooms were higher-performing classrooms or 'honors' (n=164), and 10 classrooms were regular-level classrooms or 'labs' (n=153). Prior to data collection, the six participating teachers implemented the district's eight-week genetics unit covering typical high school level genetics concepts in classical and molecular genetics.

As described above the 31 OMC items were designed to gauge students' understanding of five constructs in the genetics LP. Most of our items included response options that mapped onto 2-4 levels of a particular construct. Overall, at least 3 items, and most often 5 items, measured each level in each construct. Assessments were administered in two comparable forms with the same set of items differently ordered on each form. Across all constructs, the response options mapped onto Levels 0 through 3 of the genetics LP, as well as Level '-' which refers to distractors unrelated to any specific level on the LP. Table 1 shows the actual distribution of the items across constructs and levels, for items that had valid level-scored data. For example, although in the assessment we had more than 3 items that were designed to measure Level 3 for construct B, from students' actual responses we had Level 3 answers for only one item among the three construct B items.

Table 1. Actual number of items across constructs and levels of the genetics LP

	Level – (irrelevant)	Level 0	Level 1	Level 2	Level 3
Construct A	1	2	7	4	3
Construct B	1	5	4	4	1
Construct C	4	8	10	8	6
Construct E	0	0	5	4	3
Construct F	3	4	12	10	7

Analysis

We performed MIRM and LCA analyses separately and consecutively to answer the research questions. The primary goal of the MIRM was to estimate the difficulties of individual items and the abilities of students on the same scale. We took a confirmatory approach that assumes ordered response categories and correlations between constructs. The model also assumes that each of the items measures one of the five constructs in genetics LP. Within each construct, the responses to items are independent and have a Bernoulli distribution.

We used multidimensional random coefficient multinomial logit (MRCML) model (Adams, Wilson, & Wang, 1997) for polytomous data to estimate model parameters. Three types of results are provided by MIRM analyses. First, on the 'item-side', we estimate the difficulties of individual items. In particular, we calculated Thurstonian thresholds for each level of each item. Note that the number of thresholds is one minus the number of levels that the item is measuring. Next, on the 'student-side', for each individual student we estimated the abilities on five correlated constructs. Also, the correlations and covariances among the constructs are estimated. Finally, we calculated other relevant statistics such as deviance, EAP reliability, separation reliability, etc. MIRM analysis was performed using ACER ConQuest IRT software (Wu, Adams, Wilson & Haldane, 2007). ConQuest uses conditional maximum likelihood algorithm to estimate model parameters. The EM algorithm was terminated at the convergence criteria of 0.01 after 16 iterations.

The primary goal of the LCA was to examine whether the students can be placed into multiple latent groups or classes. Note that in using the latent class models we take an exploratory approach that does not assume ordered classes. That is, the resulting classes do not necessarily match the order of levels of the constructs. However, because the response categories were scored following the ordered levels of each construct, we can take a confirmatory perspective in examining if the response patterns reveal higher or lower level performances in some classes. i.e. do the classes differ in ability. We used latent class models (Lazarsfeld & Henry, 1968) for polytomous data to determine classes of students. The varying number of classes was incorporated in the model assumptions, which can be tested by comparing posterior fit statistics. The model assumes that within each class, the items are independent and have a Bernoulli distribution. Given the distributional assumptions of the items, we can express the likelihood of any set of occurrences.

Three types of results are provided by latent class analysis. First, we estimated the probability πick that the response for each item, answered by students from each of the specified number of classes, is equal to a certain response category. Next, we estimated the posterior probability that each of the students falls into each class. For each student, the sum of these probabilities across classes equals one. Finally, we estimate the posterior probability that each student belongs to each class. Latent class analysis for polytomous outcome variables was performed using poLCA (Linzer & Lewis, 2011), a software package implemented in the R statistical computing environment. poLCA uses the EM algorithm to estimate model parameters. The known problem of the EM algorithm is that a local maximum of the log-likelihood function can be found depending on the initial values. To avoid local maxima problems, we ran poLCA 100 times for each model to ensure the results are based on the model with the global maximum likelihood. We selected the results that occurred more than 65 times out of 100 runs. Since one statistic is never a perfect measure of model fit, we looked at three statistics to assess the model fits of the global solution. The first- Log likelihood is a function of the observed responses for each student and the model parameters. The second- AIC is a measure of goodness of fit of a model that considers the number of model parameters; and the third- BIC is a measure that considers not only the number of model parameters but also the sample size. Preferred models are those that minimize values of the BIC and/or AIC. We also looked at Pearson's chi-squared (X^2) goodness of fit and likelihood ratio chi-squared (G^2) statistics for the observed versus predicted cell counts. Larger values of X^2 and G^2 indicate that the particular model fits the data better.

Results

Examining LP Level Dynamics Within and Between Constructs using MIRM Results

As noted earlier we began with the MIRM 'item-side' analyses to obtain item difficulty estimates and Thurstonian thresholds for item level scores. In Figure 1, the Wright Map depicts the core advantage of MIRM approach- inference about student performance, item difficulties, and levels of LPs can all be made on the same logit scale. On the left five panels, the estimated distributions of student abilities for the five constructs are shown as bell curves. On the right five panels, the estimated thresholds of each level score of each item is shown with colored dots: level 1 thresholds with red, level 2 with green, and level 3 with blue. The gray columns indicate items, and colored horizontal lines indicate average thresholds for levels within constructs.

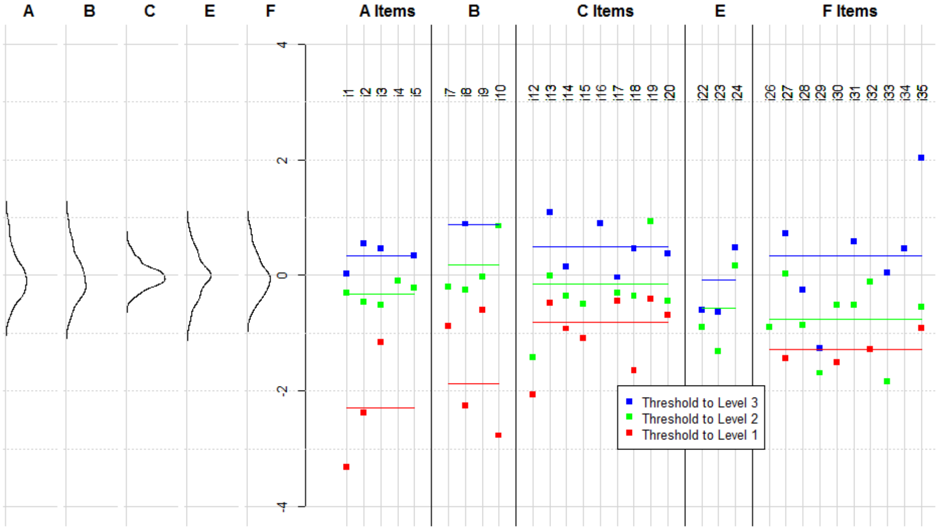

Figure 1. The Wright Map for genetics learning progression.

There are several interesting inferences that can be made from this Wright map. First, there are differences in the level thresholds across constructs. The lowest level 1 threshold across all constructs was for construct A. This lower threshold value on the logit scale implies that it was easier for students to show understanding at level 1 or above on construct A, compared to other constructs. Second, some constructs are overall easier than others. For example, construct E has the lowest threshold for both levels 2 and 3, and appears to be the easiest construct; whereas construct B appears to be the most difficult to master. Third, there are differences in the spread of level thresholds across constructs. For example, constructs A and B have a larger spread compared to C and F, suggesting that there is a larger difference between the understandings at each level of construct B. The demarcation between levels is greater for construct B compared to most of the others. Comparing level threshold spreads within a construct suggests that some 'jumps' from one level to the next are harder than others. In construct A, the difference between level 1 and levels 2-3 was the greatest. Thus while attaining a level 1 understanding on construct A is relatively easy, moving up to a level 2 understanding is much harder compared to similar moves for the other construct.

In reviewing the other side of the Wright map, the 'student-side' results of MIRM afford less inference about the levels of learning progression within or across constructs. While we can compare estimated student ability distribution between constructs, these are not particularly nuanced. We can see, for example, that students performed more similarly to each other on construct C (tighter curve) than on other constructs. However, we cannot compare the location of the distributions between constructs because the mean of the five distributions were all fixed to zero in order to allow the mean of the item difficulties to vary in the MIRM estimation. Thus, the 'between' construct comparison relates to the overall pattern, and is not about relationships between the levels across constructs. Unless standard setting and accompanying student level diagnosis are performed ad hoc (Wilson & Draney, 2002), there seem to be fewer inferences to make about level dynamics using student results from the MIRM analyses. We next present our findings from the LCA approach regarding the relationships of levels within and across constructs.

Examining LP Level Dynamics Within and Between Constructs using LCA Results

LCA allows one to identify classes of students who reason similarly across the entire assessment or individual constructs. To identify how many different classes of students exist, we fit multiple models, from two to five classes. Our results suggest that for constructs A, B, C and E, the model with two classes fitted slightly better than models with more classes. For construct F, the three-class model fitted better than two- or four or five - class models. Given that most constructs have four levels, finding the best fit in a 2-class model was unexpected. One potential explanation is that the honors and lab students function as two distinct classes that overshadow other more subtle distinctions. We subsequently decided to take the four-class solutions for all constructs and look at the characteristics of the classes in detail.

In the four-class model we estimated the conditional probability that a student in a class responds to an item with a certain response category (i.e. at a particular level). For each student and each item, the probabilities of responding to all categories are assumed to sum to one. Figure 2 summarizes these conditional probabilities. In the left graph of Figure 2, we show the results from the four-class solution with construct A items. Each of the four panels on the graph represents a predicted class. Note that class number does not necessarily match the order of the levels in LP (i.e. class 1 is not necessarily students reasoning at a level 1 on this construct). On the X axis we have five A items, and on the Y axis, we have the conditional probability of getting a certain level score for an item. Each bar represents an item and each color represents a level score for the item. The height of each segment in the bar represents the likelihood of students in that class obtaining a particular level score for the specific item. Ideally, we expect to see is that classes are different in terms of the proportion of different colors they have (i.e., there is one dominant level for each class across the items); then one can argue that the classes reflect the characteristics of the ordered levels in LP. Here, the figure shows that Level 3 (purple) responses are dominant for class 1, Level 2 (blue) is dominant for class 3, Level 1 (green) is dominant for class 4 and Level 0 (red) is dominant for class 2. Based on actual summation of the probabilities, the most dominant level in class 1 is Level 3, for class 2 it is Level 0, for class 3 it is Level 2 and for class 4 it is Level 1. This is a relatively ‘clean’ match between levels and classes. Predicted class memberships, estimated by the modal posterior probability, show that 26.5% of students belong to class 1, 16.1% belong to class 2, 11.7% to class 3, and 45.7% of the students belong to class 4.

In comparison with the cleaner LCA analysis for construct A, the right graph of Figure 2 presents the 4-class analysis for construct F, which seems to be the messiest among the five constructs. Visually, Level 3 (purple) is dominant for class 2, Level 2 (blue) is dominant for class 2,3, Level 1 (green) is dominant for class 4, Level 0 (red) is dominant for class 1,3,4, and the unrelated Level ‘-’ (orange) is dominant for class 1. Level ‘-’ represents item response options that were simply distractors and did not map onto any specific level on the LP. The summation of probabilities reveals that most dominant level for class 1 is Level 2, for class 2 is also Level 2, for class 3 is again Level 2, and for class 4 is Level 1. Predicted class memberships show that 43.5% of students belong to class 1, 48.0% belong to class 2, 5.1% to class 3, and 3.5% of the students belong to class 4.

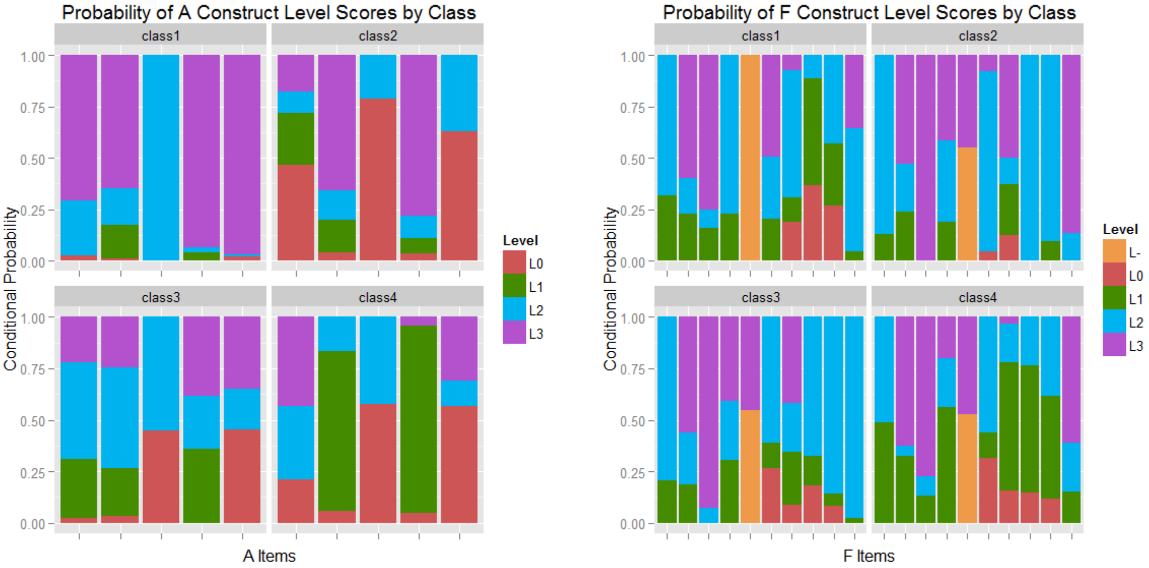

Figure 2. Estimated conditional probability of constructs A and F level scores by four latent classes.

Table 2 shows the classification of the student abilities across all five constructs, given the four-class solutions. Across the five constructs, how student abilities are classified into the four levels of the LP was clearly different. The best classification result was observed for construct A: each of the four classes in construct A matched well, also proportionally, with each of the four levels of the construct (shown by the dominant level and its proportion among all answers.) A majority of students were in class 2 (43.0%), of which student responses to the construct A items were mostly at level 2 (39.4%). However, the classification results did not match well with the four levels for all other constructs. For construct B, the four classes were characterized by only two levels of the construct: level 1 and 2. A majority of students were classified into class 3 and class 4 (48.9% + 28.4% = 77.3%), of which student responses to the construct B items were mostly at level 2 (44.9% and 54.9%). For constructs C, the four classes of students were characterized only by two dominant levels: levels 1 (class 4, 7.8%) and 2 (all three other classes, 92.2%). For construct E, the four classes were characterized by levels 1 and 3, with the majority of students at level 3 (64.4% + 21.7% = 86.1%). For constructs F, the four classes were also characterized only by levels 1 (class 4, 34.6%) and 2 (all three other classes, 65.4%).

Table 2. Classification of the students across all five constructs

Construct	Class	Predicted Class Membership	Dominant Level	Proportion of Dominant Level Answers, Across Items
A	1	0.160	Level 3	0.651
	2	0.430	Level 2	0.394
	4	0.217	Level 1	0.337
	3	0.193	Level 0	0.392
B	3	0.489	Level 2	0.449
	4	0.284	Level 2	0.549
	1	0.157	Level 1	0.628
	2	0.070	Level 1	0.383
C	1	0.312	Level 2	0.404
	2	0.525	Level 2	0.354
	3	0.085	Level 2	0.205
	4	0.078	Level 1	0.258
E	2	0.644	Level 3	0.747
	3	0.217	Level 3	0.450
	1	0.095	Level 1	0.601
	4	0.044	Level 1	0.750
F	1	0.111	Level 2	0.377
	2	0.236	Level 2	0.453
	3	0.308	Level 2	0.476
	4	0.346	Level 1	0.352

Overall, LCA, like MIRM, affords making some interesting inferences. First, as noted earlier, the best fitting model is a two-class model rather than a four-class model, suggesting that the proposed levels of the progression may not map neatly, or at all, onto students' actual performance. Second, when using a four-class model we find that some constructs are much messier than others. By this we mean that the classes in some constructs map poorly onto levels (construct F) compared with classes in other constructs (construct A). Often there are clearer class-level association for the highest and lowest performing classes (cleaner mapping onto the least and most sophisticated levels of the construct map) and a much messier middle, a phenomenon that has been previously documented (Gotwals & Songer, 2010; Steedle & Shavelson, 2009). Students in this messy middle tend to reason inconsistently, performing well on some items and less well on others. Third, in some cases one can make comparisons between classes across different constructs. For example, our analysis suggests that the students who are classified in class 3 on construct A are mostly classified in class 1 on construct B (not shown). Both these classes (class 3 in construct A and class 1 on construct B) reason at a level 3 on both constructs respectively. However, making such comparisons between constructs A and F is problematic due to the rather fuzzy distinctions between classes on construct F. Thus the messier the constructs the more difficult it is to compare them and make inferences about cross-construct relationships.

Discussion

Overall, our results suggest that, not surprisingly, the MIRM and LCA analyses together provide more detailed and nuanced information than each alone. We have shown that MIRM provides useful information about how the levels of the items within constructs and across constructs are perceived by students. That is, which items are easy and which are hard, which constructs are overall easier and which are harder. However, MIRM does not provide much useful information about how certain groups of students within our sample behaved differently, within and across constructs in attaining different levels on the items and consequently constructs. The student ability estimates are provided with an assumption of a continuous scale, not with distinct classes, groups, or levels. We can later classify the students onto levels of LPs, but it depends on 'item-side' threshold results that do not account for difference among student groups. Consequently, MIRM may not be sufficient in understanding the problem of interest: the relationships among the levels of LPs within and between constructs. Our findings suggest that LCA is useful in providing additional, 'student-side', information about the 'messy middle' levels or classes in certain constructs of LPs. However, LCA is less amenable to ranking students' performances or to assess correlation between the performances on multiple constructs. This is because ranking and correlations require continuous data, yet LCA allows classification of students onto a few distinct levels of performances.

The benefit of using multiple measurement models and approaches to studying LPs has been noted by other researchers (Briggs & Alonzo, 2012) and there are several different approaches that have been used besides the more frequently-used MIRM (e.g., attribute hierarchy method (AHM; Briggs, Alonzo, Schwab, & Wilson, 2006), Bayesian networks (West et al., 2010). In this research, we chose to bolster the popular use of MIRM in LP research with the use of LCA in order to more fully explain the relationships among the levels of LP within and between constructs. This was possible because LCA provided student-side information that matches the discrete nature of the LP levels. While LCA, AHM, and Bayesian networks can classify students into discrete classes, LCA is a less diagnostic but simpler approach. With the cost of more detailed diagnostic information, LCA does not require a-priori specification of a matrix that formally associates items and attributes, as in AHM, nor a multitude of conditional probability tables, as in Bayesian networks, Navigating between multiple methodological frameworks for empirical validation of LPs is already a problem for researchers when resources do not allow clear guidance in the pool of methodologies. While a more formal, comprehensive comparison of methods should follow to further inform researchers, this study contributes to the ongoing scholarship on LPs by providing some reasons and guidance for choosing between the MIRM and LCA approaches given a specific research goal. By doing so, this study motivates further discussions about what types of evidence each of these methodologies provides, or not , in relation to different research questions.

The work also highlights some important implications regarding the genetics progression and learning genetics more specifically. For example, there are certain ideas that are easier for students to master than others. In this case we found that reasoning about the hierarchical organization of the genetic information (connection between DNA, chromosome, genes, nucleotides) and the universal nature of the genetic information (all organisms have genetic information that is used by their cells in essentially the same way) was an idea (construct A) that was relatively easy for students to master. However, understanding what the information is about and the cell uses the information, was the hardest idea (construct B). This may be a reflection of the common instructional focus on structure and process (structure of DNA, processes of transcription and translation) rather than on the big idea that genes are instructions for proteins and that proteins are the physical mechanism that generates our traits (Duncan & Reiser, 2007). On the other hand, understanding that parents give half the genetic information to their offspring was an easier idea for students to master and the movement up the levels of this construct (E) involved much smaller conceptual jumps .

There are also interesting differences in the spread of levels of understanding for different constructs. For example, construct C (role of proteins) shows much less variation in understanding across students. Most students are at a level 2. In comparison constructs A and B show greater variation. We believe this reflects differences in the nature of the constructs and the extent to which students have substantive prior knowledge about those ideas. It seems that students may develop understandings of construct C as a result of instruction (recall the data shows were collected after genetics was taught) and thus have fairly similar "party line" understandings of proteins. However, their understandings of the nature of the genetic information (construct B) are likely informed more extensively by prior knowledge from various sources and students may exhibit such understandings on the assessment resulting in a larger spread of student ability for that construct.

In terms of using our analyses can be used to revise the progression the picture is rather fuzzy. The relatively small sample of the study and the fact that the instruction was not based on the expectations of the progression makes drawing clear-cut conclusions difficult. The point about instruction is rather critical. If we assume that students will progress along a hypothetical progression when they experience instruction that supports such progression (i.e. instruction that capitalizes on the developmental constraints and affordances embodied in the progression), then the nature of the instruction experienced becomes a critical part of the equations. If instruction is not designed based on the progression it is not clear whether the expectation for anticipated student progress should hold.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.

Alonzo, A. C, & Gotwals, A (Eds.) (2012). *Learning progressions in science: Current challenges and future directions.* Rotterdam: Sense Publishers.

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education, 93*(3), 389–421.

Briggs, D. C., & Alonzo, A. C. (2012). The psychometric modeling of ordered multiple-choice item responses for diagnostic assessment with a learning progression. *Learning Progressions in Science*, 293-316.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. R. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*(1), 33-63.

Brown, N. J., Nagashima, S. O., Fu, A., Timms, M., & Wilson, M. R. (2010). A framework for analyzing scientific reasoning in assessments. *Educational Assessment*, *15*(3-4), 142-174.

Duncan, R. G., & Hmelo-Silver, C. (2009a). Learning progressions: Aligning curriculum, instruction and assessment. *Journal of Research inScience Teaching, 46*(6), 606–609.

Duncan, R. G., Rogat, A., & Yarden, A. (2009b). A learning progression for deepening students' understanding of modern genetics across the 5th-12th grades. *Journal of Research in Science Teaching, 46*(6), 644-674.

Hadenfeldt, J. C., Neumann, K., & Liu, X. (2013). Validating a model of students' progress in understanding the concept of matter. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Rio Grande, Puerto Rico.

Gotwals, A. W., & Songer, N. B. (2010), Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education, 94*(2) 259–281

Jeon, M., & Rabe-Hesketh, S. (2012). Profile-likelihood approach for estimating generalized linear mixed models with factor structures. *Journal of Educational and Behavioral Statistics, 37*(4), 518-542.

Jordan, R. & Duncan, R. G. (2009). Student teachers' images of science in ecology and genetics. *Journal of Biological Education, 43*(2), 62-69.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Houghton, Mifflin.

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software, 42*(10), 1-29.

Mohan, L., Chen, J. and Anderson, C. W. (2009), Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching, 46*(6), 675–698.

National Research Council [NRC]. (2007). *Taking Science to School: Learning and Teaching Science in Grade K-8.* Washington DC: The National Academy Press.

Plummer, J.D. and Krajcik, J.S. (2010). *Building a Learning Progression for Celestial Motion: Elementary Levels from an Earth-Based Perspective. Journal of Research in Science Teaching*, 47(7): 768-787

Rivet, A., & Kastens, K. (2012). Developing a construct-based assessment to examine students' analogical reasoning around physical models in earth science. *Journal of Research in Science Teaching, 49*(6), 713-743.

Shavelson, R., Stanford Educational Assessment Laboratory (SEAL) and Curriculum Research & Development Group (CRDG). (2005). *Embedding assessments in the FAST curriculum: The romance between curriculum and assessment.* Final Report.

Shavelson, R. J. (2009). Reflections on learning progressions. Paper presented at the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA.

Shea, N. A., & Duncan, R. G. (2013). From theory to data: The process of refining learning progressions. *Journal of the Learning Sciences, 22*(1), 7-32.

Steedle, J. T., & Shavelson, R. J. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching, 46*(6), 699-715.

West, P., Rutstein, D. W., Mislevy, R. J., Liu, J., Levy, R., DiCerbo, K. E., ... & Behrens, J. (2009). A Bayes net approach to modeling learning progressions and task performances. Paper presented at the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA

Wilson, M. (2013). Constructing measures: An item response modeling approach. Routledge Academic.

Wilson, M., & Draney, K. (2002). A technique for setting standards and maintaining them over time. *Measurement and multivariate analysis*, 325-332.

Wu, M. L., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne: Educational Measurement Solutions.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, (2007). ACER ConQuest 2.0 [computer program]. Melbourne: ACER.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No.1053953. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.