Using Test Responses to Validate a Learning Map of Integer Understanding

Angela Broaddus

Anu Sharma

The University of Kansas

Seth Adjei

Worcester Polytechnic Institute

Contact Information:

Angela Broaddus, PhD
Curriculum and Assessment Specialist, Senior
Center for Educational Testing and Evaluation
JR Pearson Hall, Room 608
1122 West Campus Road
Lawrence, KS 66045
785-864-2916
broaddus@ku.edu

Abstract

This study examined the validity of a learning map related to integer understanding by evaluating student test responses. A total of 2,846 middle school students responded to an assessment consisting of 25 items developed to measure proficiency on 16 integer skills in the learning map. Test responses were analyzed using Bayesian analytic techniques and cross validation to evaluate model fit. Qualitative measures including item alignment studies and implications from relevant mathematics education literature informed revisions to the learning map. Benefits of the mixed methods approach, limitations, and implications for future research are discussed.

Using Test Responses to Validate a Learning Map of Integer Understanding

Introduction and Background

Effective mathematics teaching requires teachers to have deep understanding of the content they are responsible to teach and to understand how students construct mathematical knowledge (NCTM, 2014). To this end, the Teaching and Learning principle establishes learning progressions as effective tools for planning activities, creating materials, and developing tasks and assessments. To support these uses, however, these tools must be analytically refined using data about student learning collected within real-world settings (Shea & Duncan, 2013).

Learning progressions describe how understanding of particular content develops over time and experience (Common Core Standards Writing Team, 2013; Popham, 2011). Learning trajectories (Clements & Sarama, 2004) and learning hierarchies (Gagné, 1968) also describe learning sequences, articulating pathways by which students construct new understandings by connecting prior knowledge to new ideas. In comparison, and adding to this array of models of student learning, a learning map provides a network representation of learning that encompasses linear and hierarchical models, indicating prerequisite relationships between learning targets and permitting multiple pathways when appropriate for accessibility by different types of learners. In this study, we describe how a section of a learning map was evaluated and refined using both qualitative and quantitative techniques.

Learning Map Development

Since 2011, researchers at a large Midwestern university have been developing a comprehensive learning map (LM) of the mathematics children are expected to learn from birth through high school. A learning map consists of nodes and connections; each node describes a concept or skill, and connections indicate prerequisite relationships among nodes. Nodes and connections in the LM were defined by consulting relevant mathematics education and cognitive psychology literature, which produced an LM containing 2,579 nodes and 5,360 connections, suggesting a fine-grained view of mathematical learning.

While initial map development was primarily grant-funded as the basis of an assessment program, we propose that the LM can be used by educators to improve instruction and learning for all students. We believe the LM can be used by teachers, who may lack deep understanding of mathematics they are responsible to teach, to deepen their mathematical knowledge, analyze standards and curriculum, and design productive instruction. However, the potential utility of the LM to guide instructional planning or assessment development is limited by its accuracy as a representation of student learning. This study responds to the call to develop systematic approaches for revising learning models using empirical data (Shea & Duncan, 2013), with our focus being on a LM section depicting the development of proficiency with integers.

Validation Techniques

Learning models can be validated using qualitative techniques (e.g, student interviews and instructional observations) or quantitative methods, (e.g., test responses) (Leighton, Gierl, & Hunka, 2004). Multiple statistical models are available to analyze the fit of data linked to a theoretical learning model to that learning model, while other methods use response data solely to determine a learning model. Wang (2005) proposed a genetic algorithm based method for determining optimal curriculum for schools, which reduced the amount of time needed to arrange academic courses into an optimal curriculum. Cen, Koedinger, and Junker (2006) described a process for analyzing multi-dimensional skill maps (e.g., the LM), whereby successive adjustments to a map were analyzed to determine the arrangement of nodes and connections that best fit available data. Desmarais, et. al. (2007) presented a framework for identifying structures from student data and called these structures Partial Order Knowledge Structures (PoKS). The PoKS framework is probabilistic in nature and infers the structure from the student's responses to a poll of items. One limitation of this method is that when the number of items is large, the approach is not scalable. In the current study we combined learning theory with quantitative analytical techniques to evaluate a section of the LM related to integers, using data collected from student responses to a test developed for the same section of the LM.

Different techniques are available to examine how accurately a statistical model derived from data represents, or fits, the data. Models can overfit or underfit the data from which they are derived. A model that overfits falsely includes random fluctuations in the data, resulting in a model that represents the specific data very well but that does not adequately represent the meaningful relationships in the data more generally. A model that underfits fails to capture the meaningful relationships in the data. Either source of ill fit leads to faulty predictions when the model is used to generate new information. An effective method to eliminate the problem of overfitting or underfitting is to derive a model from one data set and then test that model on a different data set. A variation of this method is to use k-fold cross validation (Browne, 2000, Refaeilzadeh, et. Al. 2009), which is a method for iteratively splitting the available data k-times so that some of the data is used to generate a statistical model and the remaining data is reserved for testing the fit of that statistical model. In an initial implementation of k-fold cross validation, the data are split casewise, where some percentage of the cases are used for training the model, and the remaining percentage of cases are used to test the model.

When considering the accuracy of learning models in the presence of student test responses, one can ask, "Does the model allow you to generalize to new students?" and "Does the model allow you to generalize to new problems?" Whereas k-fold cross validation can account for each of these concerns independently, more complex implementations attempt to account for both concerns simultaneously. In such an implementation, the data are split into multiple folds along one or more dimensions. For example, in a study with 100 cases (i.e., examinees), the cases might be split into five folds of 20 cases per fold. Then a model could be trained on 4/5 of the data and tested on the remaining 1/5 fold (i.e., holdout set). Next, the model could be tested on different aspects or cases within the holdout set, in which case the data would be split again into multiple folds along the observations (i.e., item responses).

Integers

The LM section pertaining to integers was selected because (a) understanding, graphing, and operating with integers comprise important middle school mathematics topics, yet relatively few studies have investigated reasoning with integers, and (b) understanding and working with integers challenges students, who try to apply their whole number schemes to integers (Bishop et al., 2014). Specifically, students who cling to the whole number property that adding always produces larger numbers become confused when they attempt to add a positive number to a negative number. Because symbolic representations of integers can be confusing, students often experience integers initially through problem solving contexts involving assets and debts, sea level, or temperature; however, these opportunities permit students to circumvent the need to acknowledge that negative numbers possess both magnitude and direction because negative values in context can be labeled differently rather than assigned a negative sign (Peled & Carraher, 2008). Nevertheless, cases where students must explain that $5 – $7 results in a debt instead of an asset provide productive opportunities for introducing integers.

Integer notation also causes confusion because the same symbols used for addition and subtraction with whole numbers gain new meanings when used with integers. Students must expand their understanding particularly of the "–" sign to incorporate its meaning as *negative, opposite,* or *minus* (Lamb et al., 2012). Moreover, the meanings of positive and negative signs can change within a problem, requiring students to incorporate their understanding of signs with their knowledge of operations on integers.

The number line model poses additional challenges to students learning to understand and perform operations with integers. Students initially tend to separate the number line at zero (Peled, Mukhopadhay, & Resnick, 1989). They first view the number line to the left of zero as having similar rules to the number line to the right of zero, and they struggle to coordinate these *divided number lines* into a *continuous number line*. These aspects challenge students as they struggle to understand integers as numbers with magnitude and direction and as a set that is symmetric around zero.

Research Questions

The purpose of this study was to evaluate the accuracy of the LM as a representation of student learning. We describe how a LM, initially created in response to what mathematics education scholars suggest about understanding of integers, was analyzed and transformed using a combination of analytic techniques and iterative qualitative reviews of the LM nodes and test items used to generate the data. The study was guided by one overall research question and two sub-questions:

- What insight is gained about the accuracy of the LM section pertaining to integers from analysis of data generated from an assessment informed by that LM section?
  - Which nodes in the LM section are distinct and which nodes are candidates for merging?

- Which ordered connections depicted in the LM are consistent with data and which connections are reversed?

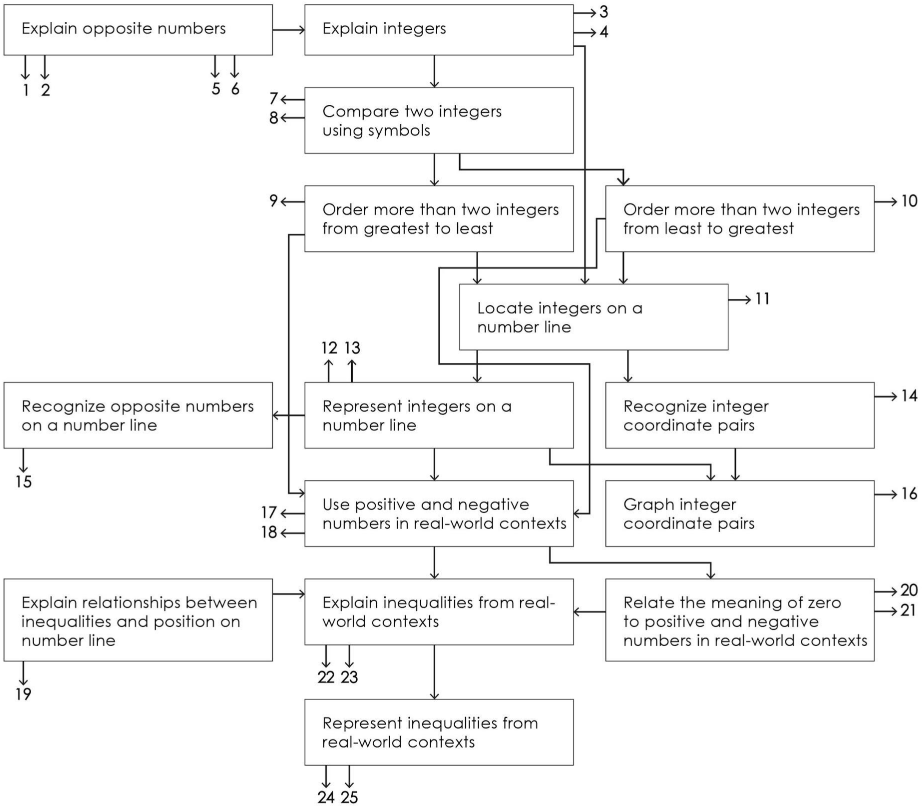
Answers to these research questions informed adjustments to the LM before providing it as an instructional resource. This study illustrates how researchers can use large-scale data sets to evaluate hierarchical or networked models of student learning such as the LM.

Method

The present study followed principles of evidence-centered design (Mislevy & Haertel, 2006), which requires test items and forms to be developed explicitly to assess specific concepts and skills (Huff & Goodman, 2007). A domain study yielded the initial student model, i.e., LM section pertaining to integers, which guided the development of test items aligned to that student model. The resulting structure shown in Figure 1 constituted a Bayesian network, where the nodes in the LM section represented latent variables, and the test items represented observable variables. It is important to note that the LM section presented in this paper only includes the latent nodes selected for test development. In reality this section is subsumed by a comprehensive learning map that includes both prerequisite and post-requisite connections to and from the nodes discussed herein.

To address our research questions, we applied a combination of quantitative and qualitative techniques. We first applied a computer algorithm to the node and item arrangement shown in Figure 1 to compare different models in terms of how accurately they predicted student responses. The model identified by the algorithm to have the best overall fit to available data is shown in Figure 2. We reviewed the recommended model resulting from the algorithmic approach, taking into consideration the nature of the test items as well as the latent structure of the nodes. Then we revisited the mathematics education literature and considered several additional adjustments to the learning map node structure, yielding the model shown in Figure 3. The next sections describe these analyses and adjustments.

Figure 1. Learning Map Section Related to Integers – Form 1


Explain opposite numbers
1 2
5 6
Explain integers
3
4
Compare two integers using symbols
7
8
Order more than two integers from greatest to least
9
Order more than two integers from least to greatest
10
Locate integers on a number line
11
12 13
Represent integers on a number line
Recognize opposite numbers on a number line
15
Recognize integer coordinate pairs
14
Use positive and negative numbers in real-world contexts
17
18
Graph integer coordinate pairs
16
Explain relationships between inequalities and position on number line
19
Explain inequalities from real-world contexts
22 23
Relate the meaning of zero to positive and negative numbers in real-world contexts
20
21
Represent inequalities from real-world contexts
24 25

Note: Rectangles with text descriptions are the LM nodes. The numerals connected to each rectangle indicate the numbered test items aligned to each LM node.

Algorithmic Approach to Model Refinement

We created and applied a *hill climbing* algorithm that began with the original structure shown in Figure 1 and successively merged nodes until the model was reduced to a single theoretical skill (i.e., node). At each stage within this process, items were realigned to their merged skills, the model was tested using cross validation, and model fit (i.e., the accuracy with which the model predicted student responses) was assessed using Root Mean Square Error

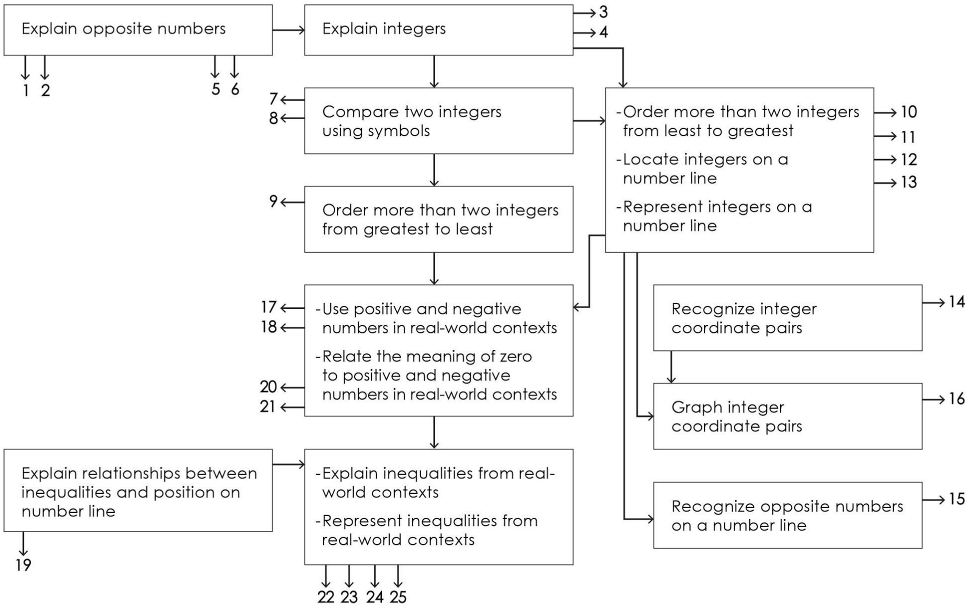
(RMSE). RMSE was calculated by squaring the difference between each actual value and predicted value and then finding the average value of the squared differences. Taking the square root of the average yielded the RMSE value for each model. Details about this process are described by Adjei et al. (2014).

Results of the *hill climbing* process suggested three node combinations shown in Figure 2. We analyzed each suggested merge by considering the skills or concepts described by the affected skills as well as the test items associated to those skills. Generally, the three merges suggested in Figure 2 indicated that the items associated with the nodes being merged did not adequately draw out the distinct skills indicated by the nodes. We suspect the multiple-choice format of the test questions to be partially responsible. For example, three nodes that appeared very similar when elicited using a multiple-choice format were *ordering more than two integers from least to greatest, locating integers on a number line,* and *representing integers on a number line.* In these cases, the test items presented sets of integers and required the student the select the answer choice that displayed the set in order from least to greatest. The use of a number line effectively worked as a particular strategy for ordering the integers. However, locating a particular integer that is correctly plotted on a number line is likely a different skill than creating a number line model and plotting a specific integer on that number line. The nodes and items associated with each recommended merge are discussed in Appendix A.

Qualitative Approach to Model Refinement

Results from our initial algorithmic approach prompted us to revisit available literature concerning children's learning to understand and operate with integers. In response to the results from the *hill climbing* process described by Adjei et al., (2014), what we learned from the literature, the alignment of the test items to nodes, and in consideration of limitations of the

Figure 2. Learning Map Section Related to Integers – Form 2


Explain opposite numbers
1
2
5
6
Explain integers
3
4
7
8
Compare two integers using symbols
-Order more than two integers from least to greatest
-Locate integers on a number line
-Represent integers on a number line
10
11
12
13
9
Order more than two integers from greatest to least
17
18
-Use positive and negative numbers in real-world contexts
-Relate the meaning of zero to positive and negative numbers in real-world contexts
20
21
Recognize integer coordinate pairs
14
Graph integer coordinate pairs
16
Explain relationships between inequalities and position on number line
19
-Explain inequalities from real-world contexts
-Represent inequalities from real-world contexts
22
23
24
25
Recognize opposite numbers on a number line
15

multiple-choice items administered to students, we adjusted the hypothesized structure. Changes included editing a node name, repositioning a node, removing three nodes, adding four nodes, and merging two pairs of nodes. After adjusting the map structure with the changes listed below, we realigned all items to the new structure, as shown in Figure 3. The implemented changes are listed in groups and are followed by explanations of the considerations that led to each type of change.

- Reduced cognitive meaning of *explain opposite numbers* to *recognize opposite numbers*
- Repositioned *explain integers* to lie later in this map section

Peled and Carraher (2006) and Kent (2000) describe children initially using counting strategies and intuition to recognize values less than zero and operate within problem situations involving debt or temperature. Such contexts provide rich opportunities for students to ground their initial experiences with negative numbers or quantities in familiar settings. Through deliberately designed instructional sequences (e.g., Gregg & Gregg, 2007) students build more sophisticated understanding, culminating in the ability to explain integers in terms of the magnitude of what they measure and their direction in relation to zero (i.e., positivity or negativity). In response to these recommendations, we re-conceptualized the two nodes to better distinguish students' initial ability to *recognize opposite numbers* from their ability to *explain integers,* and repositioned the latter node to better model it later in the LM as a culmination of understandings about integers.

- Removed *represent integers on a number line*
- Removed *explain relationship between inequalities and position on the number line*
- Removed *recognize integer coordinate pairs*
- Added *locate whole numbers on a number line*
- Added *locate negative numbers on a number line*
- Added *add two integers with different signs*
- Added *subtract two integers with different signs*

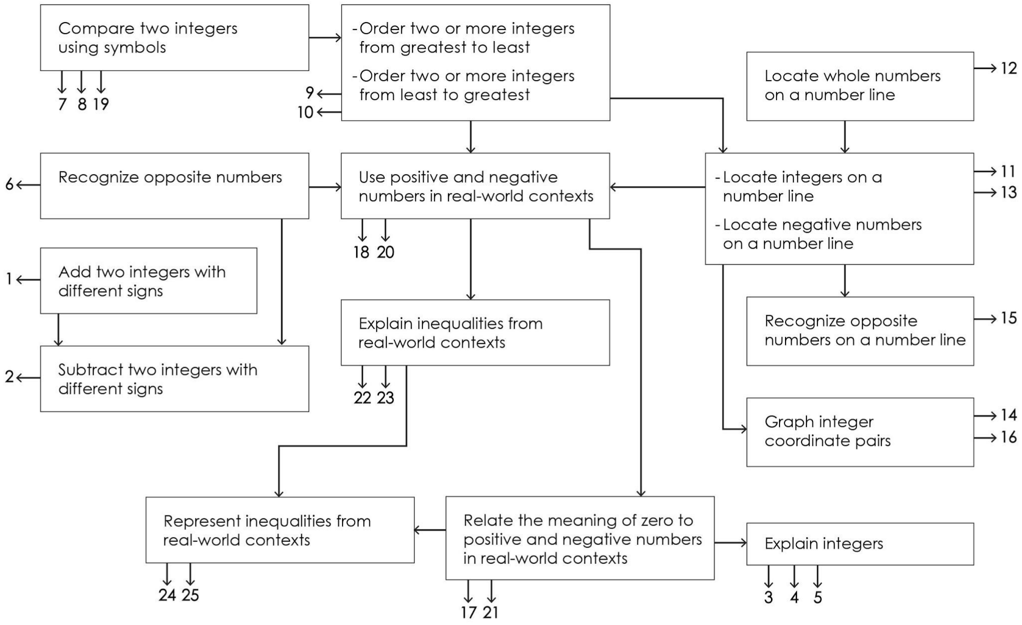
Although the items on our integers assessment were developed specifically for the nodes shown in Figure 1 and described in Table 1, our initial analyses suggested different item-node alignments. After re-examining the items, the nodes, and the literature, we determined that some of the items better aligned to skills that were not included in Figure 1 or Figure 2. Specifically, we created new nodes to disentangle *adding two integers with different signs* and *subtracting two integers with different signs* and aligned to these nodes two items that were previously associated with *explaining opposite numbers,* which better reflected the levels of integer knowledge described by Peled (1991) and Peled, Mukhopadhyay, and Resnick (1989). Similarly, and responding to these same authors' recommendations, we inserted nodes to separate the abilities to *locate whole numbers on a number line* and *locate negative numbers on a number line,* and realigned items previously associated with *representing integers on a number line* to these more specific nodes. This realignment also improved the descriptive accuracy of the nodes for these items. Whereas our integer assessment was entirely multiple-choice, no item actually required students to create their own representation of a number line. Thus our items that asked students to identify one of four number line graphs that showed a given integer graphed correctly better reflected the ability to *locate whole numbers, negative numbers, or integers on a number line.*

- Merged *locate negative numbers on a number line* with *locate integers on a number line.*
- Merged *order two or more integers from least to greatest* and *order two or more integers from greatest to least.*

Our initial data analyses indicated that these pairs of nodes were not distinguishable in our data. Because we were seeking a best fitting model, we merged these pairs of nodes, despite recommendations (e.g., Peled, 1991) to distinguish the abilities to *locate negative numbers on a number line* and *locate integers on a number line.* Discussion of this decision follows in our results section.

Each of the three node and item arrangements were analyzed to identify the model that was (a) consistent with recommendations from the literature describing student learning to operate with integers, and (b) possessed the highest fit statistics. For each model, we implemented cross validation and the model fit analyses to arrive at metrics that would help us identify the node and item arrangement with the best fit.

Figure 3. Learning Map Section Related to Integers – Form 3


Compare two integers using symbols
7 8 19
-Order two or more integers from greatest to least
-Order two or more integers from least to greatest
9
10
Locate whole numbers on a number line
12
6
Recognize opposite numbers
Use positive and negative numbers in real-world contexts
18 20
-Locate integers on a number line
-Locate negative numbers on a number line
11
13
1
Add two integers with different signs
2
Subtract two integers with different signs
Explain inequalities from real-world contexts
22 23
Recognize opposite numbers on a number line
15
Graph integer coordinate pairs
14
16
Represent inequalities from real-world contexts
24 25
Relate the meaning of zero to positive and negative numbers in real-world contexts
17 21
Explain integers
3 4 5

Model Fit Analyses

We represented each model as a Bayesian Network and used Murphy's Bayes Net toolkit for MATLAB (Murphy, 2001) to fit the models. Each representation of the models had latent and observed nodes. The latent nodes represented the skills in the learning map and the observed nodes represented the test items. An estimate of each student's knowledge of each skill was learned from the data, and this estimate was then used to predict that student's performance on other held-out items aligned to the same skill. This estimate of the students' knowledge was dependent on the estimates of his/her knowledge of the prerequisite skills of the given skill, according to the structure proposed in each version of learning map. For example, suppose skill A had prerequisite skills B and C. Our estimate of a student's knowledge of skill A was based on that student's knowledge of skills B and C. Expectation Maximization (Moon, 1996), a hill-climbing data mining algorithm, was used to learn the estimates of students' knowledge from the data set, and these estimates were then used to predict student performance on the items attached to the nodes in the learning map.

Cross Validation

To evaluate the fit of each model, we investigated the accuracy of predicted responses on some items, given the responses to other items. We used cross-validation (Browne, 2000) as a means to make visible whether each model overfit or underfit the available data. In this study, we divided the data into five student folds and three item folds, where student folds were selected randomly, and item folds remained constant for each step in the analysis. For each of the five student folds, the model was trained on the other four folds of students and tested on the fifth fold that was held out during the training process. For each test set, we split the item responses into three different folds. Two of the item folds were provided to the model as evidence while the model predicted the item responses for the third fold. This process was used for each round of the cross-validation. In general the cross-validation process uses the entire test data for prediction; however in this study, we provided some of the test data as evidence to the model and tested the model on the remaining data. We applied this more complicated strategy in order to improve the predictive accuracy of the model simultaneously for both students and items. By using this complex validation process, we were able to investigate the power of the model to generalize to both unseen students and unseen items, as evidenced by the accuracy of predicting responses of specific students to specific test items. The resulting model allowed us to state how accurately we could predict new student responses.

We used Root Mean Squared Error (RMSE) of the predicted responses to measure the accuracy of each model, where a smaller RMSE indicated better fit. RMSE was used because it penalizes errors in predictions; thus a model with a smaller RMSE score had higher predictive power and was a better representation of student knowledge.

Data Sources

Data was collected in Spring 2013, from 2,846 middle school students attending public schools in a large Midwestern state. Students, whose teachers elected to administer this optional assessment, responded to 25 items developed specifically to address the16 nodes shown in Figure 1. Each skill was assessed by one or more multiple-choice items. Learning map nodes and the numbered items aligned to each node are shown in Table 1, where Form 1 indicates the item-node alignments shown in Figure 1, Form 2 indicates the item-node alignments shown in Figure 2, and Form 3 indicates the item-node alignments shown in Figure 3.

Table 1. Node and Item Alignment Information

Node Name	Form 1 Item Numbers	Form 2 Item Numbers	Form 3 Item Numbers
add two integers with different signs	NA	NA	1
compare two integers using symbols	7, 8	7, 8	7, 8, 19
explain inequalities from real-world contexts	22, 23	NA	22, 23
explain inequalities from real-world contexts; represent inequalities from real-world contexts	NA	22, 23, 24,25	NA
explain integers	3, 4	3, 4	3, 4, 5
explain opposite numbers	1, 2, 5, 6	1, 2, 5, 6	NA
explain relationships between inequalities and position on the number line	19	19	NA
graph integer coordinate pairs	16	16	14, 16
locate integers on a number line	11	NA	11
locate integers on a number line; locate negative numbers on a number line	NA	NA	11, 13
locate whole numbers on a number line	NA	NA	12

Node Name	Form 1 Item Numbers	Form 2 Item Numbers	Form 3 Item Numbers
order more than two integers from greatest to least	9	9	9
order more than two integers from greatest to least; order more than two integers from least to greatest	NA	NA	9, 10
order more than two integers from least to greatest	10	NA	NA
order more than two integers from least to greatest; locate integers on a number line; represent integers on a number line	NA	10, 11, 12, 13	NA
recognize opposite numbers	NA	NA	6
recognize integer coordinate pairs	14	14	NA
recognize opposite numbers on a number line	15	15	15
relate the meaning of 0 to positive and negative numbers in real-world contexts	20, 21	NA	17, 21
represent inequalities from real-world contexts	24, 25	NA	24, 25
represent integers on a number line	12, 13	NA	NA
subtract two integers with different signs	NA	NA	2
use positive and negative numbers in real-world contexts	17, 18	NA	18, 20
use positive and negative numbers in real-world contexts; relate the meaning of zero to positive and negative numbers in real-world contexts	NA	17, 18, 20, 21	NA

Table 2 lists classical item statistics for the items on the integers assessment. Mean item *p*-value was 0.75 with a range from 0.35 to 0.96, and we observed an inverse relationship between *p*-value and standard deviation, i.e., lower *p*-values had higher standard deviations and higher *p*-values had lower standard deviations.

Table 2. Classical Item Statistics

Item Number	P-Value	Standard Deviation	Item Number	P-Value	Standard Deviation
1	0.36	0.48	14	0.82	0.39
2	0.35	0.48	15	0.95	0.21
3	0.74	0.44	16	0.78	0.41
4	0.58	0.49	17	0.93	0.25
5	0.40	0.49	18	0.89	0.32
6	0.93	0.26	19	0.85	0.36
7	0.70	0.46	20	0.96	0.20
8	0.67	0.47	21	0.66	0.48
9	0.83	0.38	22	0.86	0.35
10	0.90	0.30	23	0.81	0.39
11	0.91	0.29	24	0.64	0.48
12	0.84	0.36	25	0.44	0.50
13	0.85	0.36	AVERAGE	0.75	0.38

Results

In this section we present the performance of each model and then describe selected results for particular items and skills. This description will show which of the components contributed to the goodness of fit, focusing on the model shown in Figure 3 as the result of our approach integrating qualitative decisions stemming from relevant literature and quantitative results from data fitting analyses.

On completion of the data fitting process, we determined the goodness of fit for each of the models. Goodness of fit was determined using four metrics: Area Under the Receiver Operating Characteristic Curve (AUC), Root Mean Square Error (RMSE), Accuracy, and R^2.

RMSE is an error metric with range [0, 1]. AUC values indicate how well a model predicts observations and has a range of [0, 1], where AUC = 1 indicates perfect fit. The RMSE measures the distance between the predictions made by model when compared to the data from which that model is derived. RMSE values close to zero indicate high accuracy rates, that is, the smaller its value, the better the model fits the data. Accuracy is another measure of goodness of fit. It measures how close a model's predictions are to the actual values in the test set. Higher values of accuracy indicate more accurate predictions and hence a better model.

Table 2. Overall Model Level Statistics

Model	AUC	RMSE	Accuracy	R^2
Figure 1	0.78003	0.38791	0.78744	0.20556
Figure 2	0.79514	0.38352	0.79093	0.22343
Figure 3	0.78771	0.38500	0.79076	0.21744

Items and Skills

A model's goodness of fit is affected by both the hypothesized relationships among the LM nodes and the alignment of the test items to those nodes. In other words, when items are not associated with the correct nodes (e.g., they draw on different skills than what their nodes describe), then the items contribute negatively to the goodness of fit of the model. In order to investigate the effect each of the items had on how well the models predicted students' responses, we analyzed results at the item level to determine the individual statistics for each item. Table 3 lists items that had high accuracy and low RMSE values for all three models, which suggests that these items were properly aligned to nodes and indicates that the region of the model containing these items fit the data well. Similarly, Table 4 lists the nodes that contributed positively to model fit.

Table 3. Item level results

Item Number	Mean RMSE (SD)	Accuracy (SD)
20	0.194 (0.001)	0.961 (0)
15	0.210(0.002)	0.954(0)
17	0.247(0.001)	0.934(0)
6	0.261(0.000)	0.927(0)
11	0.289(0.002)	0.907(0)
10	0.293(0.002)	0.901(0)
18	0.313(0.001)	0.888(0)
22	0.333(0.005)	0.856(0.002)
13	0.354(0.003)	0.852(0.000)

Table 4. Node Level Results

Node	Model	RMSE
recognize opposite numbers on a number line	Figure 1 Figure 2 Figure 3	0.20496 0.20964 0.21058
recognize opposite numbers	Figure 1 Figure 2 Figure 3	NA NA 0.26082
use positive and negative numbers in real-world contexts	Figure 1 Figure 2 Figure 3	0.28239 NA 0.25914

These results suggest a few skills that performed well at predicting student responses. In particular, *recognizing opposite numbers on a number line* and *using positive and negative numbers in real-world contexts* had a consistently low RMSE for the two models. After the node *recognize opposite numbers* was re-characterized as described earlier, its predictive power improved, indicating that our combination of evaluating analyses and reflecting on the literature led to valid model refinement decisions.

Some nodes consistently performed poorly. Two nodes with consistently poor RMSE (e.g., RMSE ≥ 0.45) were *explaining* and *representing inequalities from real-world contexts*. This poorness of fit was identified in all models, suggesting that these skills may be out of place and require further scrutiny to ascertain their placement in the learning map.

Discussion

In our exploration of model fit to a learning map section based on data collected from an assessment containing all multiple choice test items we incorporated recommendations from the literature to evaluate the meaning of analytic results. We found that our initial learning map section did not adequately describe the fine-grained steps in learning students have been observed to take as they develop sophisticated understanding of integer concepts and operations. In response, we modified the learning map and realigned test items based on the skills required to answer them correctly and the behavioral expectations for each item (e.g., locate vs. create a representation).

Our results provide insight into the absolute necessity of item alignment to node descriptions and the need for multiple test items to elicit specific skills. One problem we observed was that in the presence of a fine-grained learning map, such as the one used in this study, two items intended to draw on different skills appeared to draw on one skill, as was the case for item numbers 11 and 13. Where item 11 included graphs with both positive and negative integers, and was intended to draw on the ability to *locate integers on a number line,* item 13 only included graphs of negative integers, and was intended to draw on the ability to *locate negative numbers on a number line.* Our model fit improved when we merged these two nodes, yet this merge is inconsistent with the literature describing students operating on the divided number line before they can operate effectively on the continuous number line (Peled, Mukhopadhyay, & Resnick, 1989). We suspect that if we used more items for each of these nodes and collected data from students with a wider range of knowledge of integers, these nodes would appear to be more distinct.

Alternatively, we observed other pairs of items, intended to draw on separate skills, that more convincingly drew on one skill, as was the case for item numbers 9 and 10. Where item 9 required students to *order integers from greatest to least*, item 10 required students to *order integers from least to greatest*. We did not identify recommendations to separate these skills, and our analyses indicated that these were inseparable in our data. Thus this merge of nodes was supported both qualitatively and quantitatively, and yielded improved model fit.

We believe our integrated, and cross-disciplinary approach to model refinement contributes to the literature describing validation of learning progressions, yet we acknowledge that the generalizability of our results may be limited by aspects of our data and features of our learning map. Specifically, we must endeavor to create and administer multiple items per node and collect data from students with a wide range of knowledge and instructional experiences. Future work should investigate whether the same patterns in the data suggest similar learning sequences among students from different demographic groups. The field should also investigate systematic ways of using qualitative data (e.g., student observations) collected in classrooms to refine and validate learning models.

Conclusions

This study's focus was to evaluate the accuracy of a LM section pertaining to integers in terms of the distinctness of the latent nodes and the order of the connections between latent nodes. Because the data was collected from test responses, we felt it necessary to consider our results by reviewing both the latent structures (i.e., nodes and ordered connections) and the test items, with particular interest in each item's alignment to the intended latent node. Our process included both quantitative analyses of goodness of fit and qualitative judgments about the skills evoked by each test item, the nature of the skills described by the learning map nodes, the alignment of test items to learning map nodes, and the relationships among the nodes in the learning map. We relied on mathematics education literature to guide and shape our interpretations and decisions.

This study provides a description of how statistical analyses can identify aspects of learning theories that do and do not resound with data collected from actual students as well as how to consult relevant literature when considering whether and how a learning theory may need to be adjusted to better reflect actual student learning. The process we followed for modifying the learning map structure included decisions based on research studies investigating student learning. While recommendations from the literature provide a sound basis for constructing theories of learning, sophisticated statistical analyses are now available for testing these theories, to identify potential flaws, and improve our understanding of student learning. Using these methods, researchers can perform empirical studies to determine the strength of proposed prerequisite skill relationships and potentially identify ideal sequences for teaching.

References

Adjei, S., Selent, D., Heffernan, N., Pardos, Z., Broaddus, A., & Kingston, N. (2014, June). Refining Learning Maps with Data Fitting Techniques: Searching for Better Fitting Learning Maps. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 413–414). Available at http://educationaldatamining.org/EDM2014/uploads/procs2014/posters/89_EDM-2014-Poster.pdf

Bishop, J. P., Lamb, L. L., Phillip, R. A., Whitacre, I., Schappelle, B. P., & Lewis, J. (2014). Obstacles and affordances for integer reasoning: An analysis of children's thinking and the history of mathematics. *Journal for Research in Mathematics Education*, *45*(1), 19-61.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, 44(1), 108-132.

Cen, H., Koedinger, K. R., & Junker, B. (2006). Learning factors analysis: A general method for cognitive model evaluation and improvement. In M. Ikeda, K. D. Ashley, T. W. Chan (Eds.) Proceedings of the 8th *International Conference on Intelligent Tutoring Systems*, 164-175. Berlin: Springer-Verlag.

Clements, D. H., & Sarama, J. (2004). Learning trajectories in mathematics education. *Mathematical Thinking and Learning An International Journal, 6*, 81-89.

Common Core Standards Writing Team. (2013). Front matter for Progressions for the Common Core State Standards in Mathematics (draft). Tucson, AZ: Institute for Mathematics and Education, University of Arizona. Retrieved from http://commoncoretools.me/wpcontent/uploads/2013/07/ccss_progression_frontmatter_2013_07_30.pdf

Desmarais, M. C., Pu, X, & Blais, J. G. (2007). Partial order knowledge structures for CAT applications. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*. Retrieved [1/10/2015] from www.psych.umn.edu/psylabs/CATCentral/

Gagné, R. (1968). Learning hierarchies. *Educational Psychologist, 6*, 1-9.

Gregg, J., & Gregg, D. U. (2007). A context for integer computation. *Mathematics Teaching in the Middle School, 13*(1), 46-50

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education* (pp. 19-60). New York: Cambridge University Press.

Kent, L. M. (2000). Connecting integers to meaningful contexts. *Mathematics Teaching in the Middle School, 6*(1), 62-66.

Lamb, L. L., Bishop, J. P., Phillip, R. A., Schappelle, B. P., Whitacre, I., & Lewis, M. L. (2012). Developing symbol sense for the minus sign. *Mathematics Teaching in the Middle School, 18*(1), 5-9.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*(3), 205-237.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6-20.

Moon, T. K.(1996). “The expectation-maximization algorithm,” *IEEE Signal Processing Mag*., vol. 11, pp. 47–60, 1996

Murphy, K. P. (2001). The Bayes net toolbox for MATLAB. *Computing Science and Statistics*, vol. 33 page 2001

NCTM. (2014). *Principles to actions: Ensuring mathematical success for all.* Reston, VA: NCTM.

Peled, I. (1991). Levels of knowledge about signed numbers: Effects of age and ability. In F. Furinghetti (Ed.), *Proceedings of the 15th Conference of the International Group of the Psychology of Mathematics Education (PME).* (pp. 159 – 166). Available at http://files.eric.ed.gov/fulltext/ED413164.pdf

Peled, I., & Carraher, D. (2008). Signed numbers and algebraic thinking. In J. Kaput, D. Carraher, & M. Blanton (Eds.), *Algebra in the early grades* (pp. 303-328). New York: NCTM.

Peled, I., Mukhopadhyay, S., & Resnick, L. B. (1989). Formal and informal sources of mental models for negative numbers. Paper presented at the *13th international conference for the Psychology of Mathematics Education*, Paris, France.

Popham, W. J. (2011). *Transformative assessment in action: An inside look at applying the process.* Alexandria, VA: ASCD.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. *Encyclopedia of Database Systems*, 532–538. doi:10.1007/978-0-387-39940-9_565.

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.

Wang, Y. (2005). *A GA-based methodology to determine an optimal curriculum for schools.*

Expert Systems with Applications 28, 1, 163–174.

Appendix A

Results from Hill Climbing Study

Three nodes identified for merging represented the abilities to *locate integers on a number line*, *represent integers on a number line*, and *order integers from least to greatest*. The test items associated with these nodes required students to select lists of correctly ordered integers or identify the correct number line graph of a particular integer. These test items did not adequately distinguish between *locating* and *representing integers on a number line* because all of the items were multiple-choice, and none provided students the opportunity to construct their own number line representations of integers. Furthermore the inclusion of *ordering integers from least to greatest* with the other two nodes may have surfaced because using a number line is inherently, cognitively connected to *ordering numbers from least to greatest*.

Two nodes identified for merging represented the abilities to *use positive and negative numbers in real-world contexts* and *relate the meaning of zero to positive and negative numbers in real-world contexts*. The test items associated with these nodes required students to interpret problems involving integers and choose integer answers or verbal statements about integers. Two of the four test items included references to zero either as freezing point or sea level. These items were designed to distinguish between the two nodes, i.e., using integers and relating integers to zero. However, the relationship between zero and positive or negative numbers is so critical for understanding integers, that it is possible one cannot compare integers without considering their values in relation to zero.

Two nodes identified for merging represented the abilities to *represent inequalities from real world contexts* and *explain inequalities from real-world contexts*. The test items associated with these nodes required students to read problems and identify inequality statements that matched the problems. These items did not distinguish between two unique nodes, i.e., representing a problem or explaining a problem, as was suggested by the two nodes.