Analysis of Learning Map Structure for a Dynamic Assessment

Feng Chen, Amy K. Clark, Russell Swinburne Romine

University of Kansas

Paper presented at the 2015 annual meeting of the American Educational Research Association, Chicago, Illinois. This paper was developed under grant 84.373X100001 from the U.S. Department of Education, Office of Special Education Programs. The views expressed herein are solely those of the author(s), and no official endorsement by the U.S. Department should be inferred. Correspondence concerning this paper should be addressed to Feng Chen, Center for Educational Testing and Evaluation, University of Kansas, daniellechen@yahoo.com.

Chen, F., Clark, A. K., & Swinburne Romine, R. (2015, April). Analysis of learning map structure for a dynamic assessment. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Abstract

This study examines the structure of a learning map used for dynamic assessment in grades three through high school. Students were assessed on items measuring 307 English language arts nodes, a dramatically larger number of skills when compared with previous diagnostic assessment analyses. Student response data was used to model mastery probabilities at the node level, and these values were used to make recommendations regarding node-to-node connections in the learning map as well as node granularity. These findings were shared with content teams to serve as supporting evidence in their decision-making process for map revisions pertaining to the order and size of cognitive skills.

Objectives

Dynamic assessment makes use of an underlying map structure to present unique items and testlets matched to each individual student's knowledge, skills, and abilities. In a learning map, learning targets take the form of individual nodes, and these nodes are connected to reflect how individual skills contribute to and provide the foundation for the development of subsequent skills. Connections between nodes in the map represent causal hypotheses about the order of skill acquisition, whereby a parent skill would be acquired prior to learning the subsequent (child) skill.

In order to ensure the best possible match of items to students, the learning map underlying the assessment system should accurately specify the connections among nodes, as well as specify nodes at the appropriate level of the granularity. Each node should represent a single, distinctive skill. To the extent connections between nodes are out of order, the items presented to students may not accurately match their learning trajectories. Similarly, nodes

should measure a single learning target or skill. Should a node measure more than one skill, or a collection of nodes actually measure only a single skill, the validity of inferences made from student performance on items measuring those skills may be affected.

To this end, the current paper examines the structure of a learning map in a dynamic assessment environment. Connections among the nodes in the map, as well as the granularity of individual nodes are examined. Based on the statistical recommendations from these analyses, content experts review the findings and provide ultimate recommendations as to whether changes should be made in the learning map as a result of the statistical evidence.

Perspective(s)

With the growing emphasis on teacher accountability in K-12 education, various constituents are increasingly interested in administering assessments that can provide greater information on student outcomes. Educators are interested in being able to use the results of assessments to guide instruction, specifically pinpointing areas of mastery or weakness (Huff & Goodman, 2007; Trout & Hyde, 2006). Researchers also increasingly emphasize the need for richer reporting practices, beginning with the call from Snow & Lohman (1989).

In light of these needs in the educational community, diagnostic classification modeling (DCM) has emerged as one technique that can be used to provide specific feedback on student ability and areas for improvement. As the name suggests, diagnostic modeling provides the unique ability to "diagnose" or identify examinee strengths and weaknesses with regard to the specific cognitive processes underlying performance on an assessment (Gierl, 2007; Yang & Embretson, 2007). Student likelihood of mastery is represented by the probability of having mastered particular skills or attributes within an interconnected web of skill acquisition, with

values closer to 1.0 indicating greater likelihood of skill mastery, and values approaching 0.0 indicating non-mastery. By aggregating these values over students, recommendations can be made regarding the optimal map structure. Similarly, Bayesian Network Analysis can be used to hypothesize causal relationships among nodes in a learning map by representing the probability that a parent node precedes a child node.

Despite the increasing prevalence of research in the educational measurement and cognitive psychology academic communities employing DCM and Bayesian networks in simulation studies and diagnostic assessments of a small number of skills, few operational testing programs have made use of such statistical methods as the primary psychometric approach for analysis of assessment results and score reporting practices, particularly on a large scale. The current study addresses this gap in the research by highlighting how DCMs and Bayesian networks can be used to evaluate the structure of a learning map and provide statistical recommendations for modifications of the learning map underlying a diagnostic assessment system.

Method

The current study covers three critical analyses in the evaluation of the underlying structure of a learning map used in a dynamic assessment environment.

1. Analysis of the connections between nodes in the learning map
2. Analysis of node granularity in the learning map
3. Content review of statistical recommendations

The learning map underlying the dynamic assessment system reflects a complex arrangement of nodes that allow for diagnosis of skill mastery and areas for improvement. The

map specifically identifies learning trajectories in English language arts and mathematics beginning at a foundational level, with skills typically learned in infancy, and mapping that skill development through twelfth grade. In contrast to learning progressions sometimes used in diagnostic assessment, the use of a learning map better approximates cognitive skill acquisition by accommodating multiple and/or alternate pathways of learning or development.

The current version of the learning map used as the foundation upon which the assessment is built consists of 1,852 nodes in English language arts and 2,395 nodes in mathematics. These nodes are linked by a total of 4,951 connections in English language arts and 5,131 connections in mathematics. This study focuses on the English language arts portion of the learning map. A snapshot of the entire learning map is presented in Figure 1.

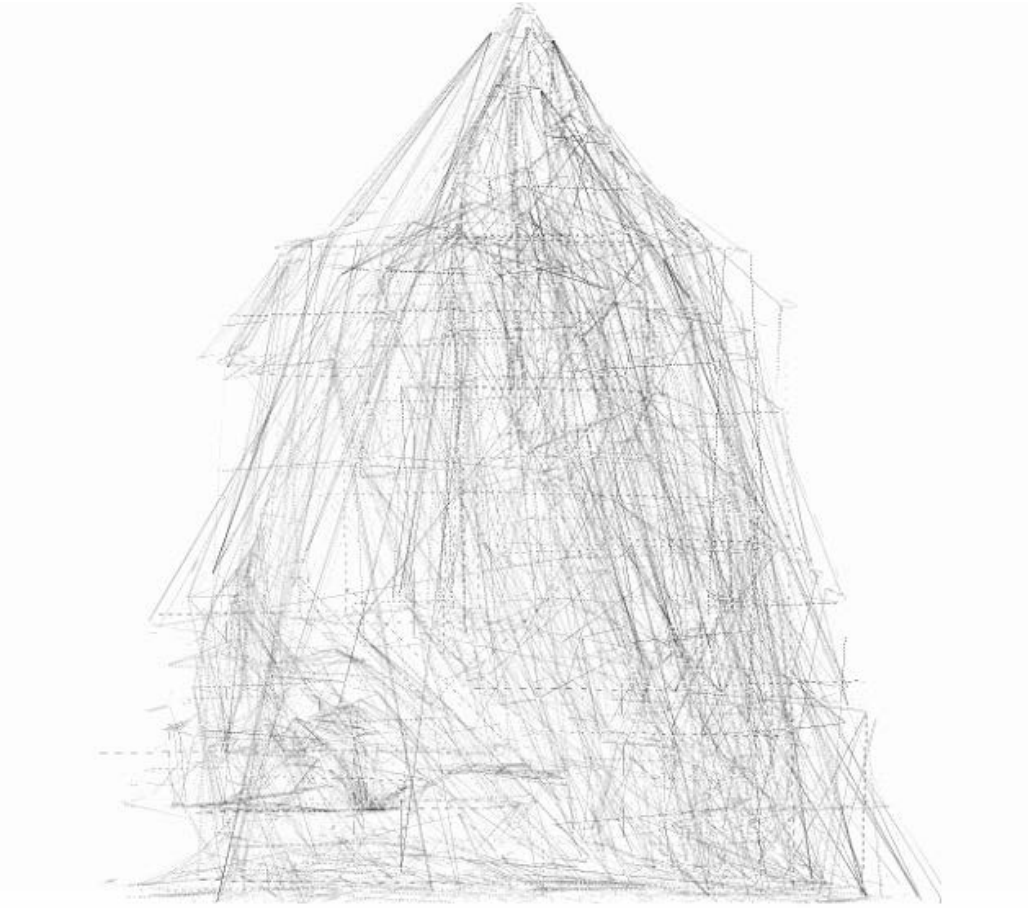

Figure 1. *Nodes contained in the learning map*

Data

According to preliminary diagnostic modeling analyses, convergence improves when sample size is greater than 100. Using a sample size of 100 as a threshold, a total of 532 English language arts testlets were included in the analysis. These testlets were administered to a total of 22,733 students in grades three through twelve across seventeen participating states between the spring of 2014 and spring of 2015. Each testlet consisted of three to eight items, and resulted in the assessment of 1,744 English language arts items. Since items were administered within testlets, a testlet effect has been accounted for in the estimation. The complete set of items

assessed allowed for data collection on 214 English language arts nodes in the learning map. A snapshot of the ELA node networks is presented in *Figure 2*.

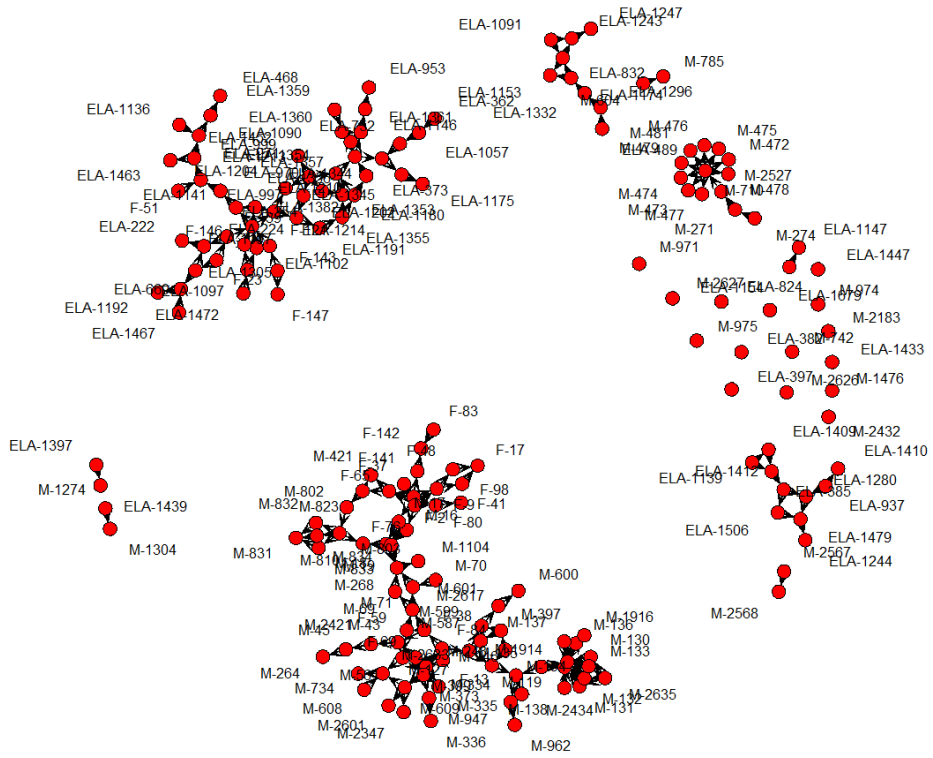

Figure 2. Nodes network diagram

Modeling

As a first step in evaluating the complex structure of the learning map, students were assessed on items written to specific nodes. The Metropolis-Hastings algorithm, a Markov Chain Monte Carlo (MCMC) method for obtaining a sequence of random samples from a probability distribution, was employed. Bayesian Inference Networks were integrated to inform causal inferences using conditional probabilities of multiple observations. All map and item parameters

were added with uniform prior distribution; all student node parameters were supplied a prior distribution driven by map parameters, and all testlet effects had a prior of a normal distribution:

$$\gamma_t \sim N\left(0, \sigma_\gamma^2\right)$$

The preliminary modeling results indicate that the model fits better with testlet effect included $DIC_{testlet}$ = 476,106.7, whereas the $DIC_{no\ testlet}$ = 518,952.1 when the testlet variance is not included. A lower number of DIC indicates that the model fits better and thus is preferred, whereby testlet effect is included in the modeling for this study.

Since the mastery status of nodes is binary, either mastery or non-mastery; items are scored as binary, either answered correctly or incorrectly, the log-odds of a correct response conditional the values of its predictors were modeled. The log-odds is also called a logit:

$$\text{logit}\left[P\left(Y_{si(t)} = 1 \middle| \alpha_{sn_1}, \gamma_t\right)\right] = \log_e\left[\frac{P\left(Y_{si(t)} = 1 \middle| \alpha_{sn_1}, \gamma_t\right)}{1 - P\left(Y_{si(t)} = 1 \middle| \alpha_{sn_1}, \gamma_t\right)}\right]$$

Here, *n* refers to node index, denoted as n_1 for node 1 and n_2 for node 2. Node status is assigned as α_{sn} = 0 for non-masters and α_{sn} = 1 for masters, with *s* as students. Testlet is denoted as *t*, and an item nested within a testlet as *i(t)*. In this case, all items have dichotomous responses where 0 = incorrect and 1 = correct. The probability function can be derived by the inverse of the logit function. For instance, for a given item in the test measuring node *a*, the probability a student *s* provides a correct response to item *i* that is nested within testlet *t* is given by:

$$P\left(Y_{si(t)} = 1 \middle| \alpha_{sn_1}, \gamma_t\right) = \frac{\exp\left(\text{logit}\left[P\left(Y_{si(t)} = 1 \middle| \alpha_{sn_1}, \gamma_t\right)\right]\right)}{1 + \exp\left(\text{logit}\left[P\left(Y_{si(t)} = 1 \middle| \alpha_{sn_1}, \gamma_t\right)\right]\right)}$$

For items measuring more than one node, analogous parameters need to be added into the model. The above log-linear probability modeling can also be converted to an intercept-slope format:

$$\text{logit}\left[P\left(Y_{si(t)} = 1 \middle| \alpha_{sn}, \gamma_t\right)\right] = \gamma_t + \lambda_{i(t),0} + \lambda_{i(t),1,n}\alpha_{sn}$$

where α_{sa} is the mastery status of student s on node a. $\lambda_{i(t),0}$ is the intercept, indicating the log-odds of non-masters providing a correct response with a testlet of average difficulty. $\lambda_{i(t),1,a}$ is the "main effect", represents the difference in log-odds of correct response between masters and non-masters of node a. γ_t is the testlet random intercept, applies to all times within a given testlet t. The probability derived from the logit function can be written as:

$$P\left(Y_{si(t)} = 1 \middle| \boldsymbol{\alpha}_s, \gamma_t\right) = \frac{\exp\left(\gamma_t + \lambda_{i(t),1,a}\alpha_{sa}\right)}{1 + \exp\left(\gamma_t + \lambda_{i(t),1,a}\alpha_{sa}\right)}$$

As shown in the formulas, the item model combines the loglinear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009) with a bifactor testlet effect. The same modeling strategy applies to both nodes and items. For instance, for a node α_{n_2} predicting a node α_{n_1}, the function is:

$$P\left(\alpha_{n_2} = 1 \middle| \alpha_{n_1}\right) = \frac{\exp\left(\mu_{n_2,0} + \mu_{n_2,1,(n_1)}\alpha_{sn_1}\right)}{1 + \exp\left(\mu_{n_2,0} + \mu_{n_2,1,(n_1)}\alpha_{sn_1}\right)}$$

Where $\mu_{n_2,0}$ denotes the intercept value of node 2, $\mu_{n_2,1,(n_1)}$ refers to the main effect of node 1 and node 2, α_{sn_1} indicates the mastery level of student *s* responding to node 1. As for node α_{n_1}, not predicted by any nodes, the probability can be described as:

$$P\left(\alpha_{n_1} = 1\right) = \frac{\exp\left(\mu_{n_1,0}\right)}{1 + \exp\left(\mu_{n_1,0}\right)}$$

Results and Discussion

The Deviance Information Criterion (DIC) for model fit was 297047.8, and the convergence rate for nodes was 90%, which is promising. Node-to-node connections in the learning map were evaluated for evidence of reversals in causal inference. A reversal node is

present when non-masters of a parent node exhibit a high chance of being a master on subsequent child nodes based on map parameters. Node reversals are detected in the map when the intercept value of the successor node is greater than zero, which leads to a probability greater than .5. In our study, 52 child nodes out of 143 total child nodes are found to have intercept values greater than zero. Of those 15 included 0.0 in the credible interval, which left 37 out of 52 child nodes, as displayed in *Table 1.* In the table, *Logit* is the log-odds of mastering the node. The *Probability* value is obtained by converting the logit to a bound of zero and one. The Heidelberger p-value is a convergence diagnostic statistic. The null hypothesis for the Heidelberger p-value is that the Markov chain is from a stationary distribution. A 5% chance the marginal posterior distribution will appear non-stationary is used as a threshold, as proposed by Heidelberger and Welch (1981; 1983). Therefore, a converged chain is desired, and a non-significant p-value is preferred.

The average probability of mastering the 37 child nodes is 0.79 (see *Table 1*), meaning that when the `parent` node is not mastered, the probability of mastering the child node is 0.79 on average. This indicates the probability of child node mastery is independent of the mastery status of the parent nodes. Conditional probability theory states a dependency exists between the children and the parent nodes, but in this case, the dependency does not hold. Because each child node has a parent node, and the estimation of a child node is conditioned on the parent node, the flagging of a particular child node results in a "reversal" in the connected parent node(s). As such, the parent nodes connecting to those 37 child nodes can be flagged as reversal nodes, leaving 43 reversal nodes in total.

Table 1. Successor nodes with intercept values greater than zero

Nodes	Logit	SD	Probability	Heidelberger P-value
ELA.1136.Intercept.1.	3.02	1.38	0.95	0.06
ELA.1147.Intercept.1.	1.44	0.18	0.81	0.17
ELA.1175.Intercept.1.	1.64	0.84	0.84	0.06
ELA.1239.Intercept.1.	0.52	0.15	0.63	1.00
ELA.1246.Intercept.1.	1.17	0.30	0.76	0.71
ELA.1248.Intercept.1.	2.12	0.99	0.89	0.15
ELA.1276.Intercept.1.	0.44	0.27	0.61	0.42
ELA.128.Intercept.1.	3.13	0.93	0.96	0.61
ELA.1339.Intercept.1.	1.00	0.14	0.73	0.23
ELA.1340.Intercept.1.	1.61	0.32	0.83	0.82
ELA.1344.Intercept.1.	2.42	0.33	0.92	0.12
ELA.1353.Intercept.1.	2.43	0.56	0.92	0.08
ELA.1356.Intercept.1.	2.07	0.97	0.89	0.05
ELA.1381.Intercept.1.	1.92	0.37	0.87	0.02*
ELA.1382.Intercept.1.	0.48	0.14	0.62	0.08
ELA.1416.Intercept.1.	2.72	0.40	0.94	0.02*
ELA.1436.Intercept.1.	1.42	0.28	0.81	0.13
ELA.1445.Intercept.1.	0.69	0.16	0.67	0.37
ELA.1461.Intercept.1.	0.47	0.19	0.61	0.43
ELA.1472.Intercept.1.	0.67	0.25	0.66	0.32
ELA.1481.Intercept.1.	0.86	0.47	0.70	0.01*
ELA.1493.Intercept.1.	0.76	0.24	0.68	0.16
ELA.1546.Intercept.1.	8.54	2.70	1.00	0.10
ELA.1550.Intercept.1.	1.27	0.42	0.78	0.97
ELA.1801.Intercept.1.	0.75	0.63	0.68	0.13
ELA.1913.Intercept.1.	1.72	0.31	0.85	0.96
ELA.2109.Intercept.1.	0.89	0.22	0.71	0.31
ELA.361.Intercept.1.	3.35	0.63	0.97	0.42
ELA.362.Intercept.1.	0.57	0.22	0.64	0.78
ELA.485.Intercept.1.	0.46	0.14	0.61	0.75
ELA.489.Intercept.1.	1.67	0.31	0.84	0.92
ELA.953.Intercept.1.	1.19	0.31	0.77	0.90
ELA.971.Intercept.1.	3.33	1.21	0.97	0.89
F.111.Intercept.1.	4.23	0.59	0.99	0.79
F.114.Intercept.1.	0.53	0.22	0.63	0.28
F.139.Intercept.1.	0.31	0.13	0.58	0.01*
F.180.Intercept.1.	2.12	0.27	0.89	0.14

**significance level p < .05*

The probability of mastering the child node should increase given the mastery of the parent nodes, as ruled by conditional probability theory. The problem with the reversal nodes is that the probability of mastery of a child node is high even when a student hasn't mastered the predicting node, which causes the connection and granularity of the nodes to be implausible.

Take node *ELA-1136* for instance. *ELA-1136* is a child of two nodes: *ELA-999* and node *ELA-1141* (see *Figure 3*).

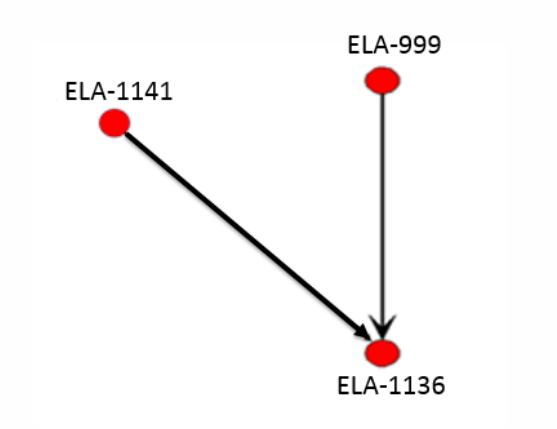

Figure 3. Example of node connection

From *Table 1*, the intercept value of *ELA-1136* is greater than zero, and thus *ELA-999* and *ELA-1141* are flagged as reversal nodes. The conditional probability of child node mastery depending on the parent node is presented in *Table 2* as an illustration. The probability of mastery of node *ELA-1136* is greater than .5, even with non-mastery of the parent node *ELA-1141*. Therefore, the connection between these two nodes is questionable, in that the causal inference is not sustained. In other words, mastery of node *ELA-1141* does not necessarily have a causal effect on the likelihood of mastering *ELA -1136*.

Table 2. Example of conditional node probability

	ELA- 1136	
ELA - 1141	Master	Non-Master
Non-Master	.96	.04
Master	.99	.01

Beyond investigating node connections for possible reversals, nodes could also potentially be overspecified. Over-specification would occur in instances where two nodes are not distinguishable from one another, or said another way, performance on the parent node perfectly predicts performance on the child node. Overspecified nodes are identified by having an intercept value less than -4, and a main effect value greater than 8. Among the 214 nodes with a sample size of at least 100, there were no overspecified nodes evident, meaning that all the examined nodes were reasonably distinct from their precursors, and did not need to be collapsed.

In addition to node level estimation, item level examination is also critical in terms of informing test construction, scoring, quality control, and other features of test development. Statistically, good items are those with low intercept and/or high main effect values, whereas non-informative items are those with high intercept and/or low main effect values. That is to say items are expected to discriminate well between masters and non-masters of the node. To flag non-informative items, we flag items with an intercept value greater than 1.0 and/or a main effect value less than 0.5. In this study, 305 out of 1,744 items were flagged for additional content expert review.

To illustrate, take items measuring node *ELA-1141* as an example. There were 68 items testing *ELA-1141*, one of which was flagged as non-informative items, as presented in *Table 3*. In theory, since *ELA-1141* is the predicting node, mastery of the node should increase the probability of answering the item correctly. Moreover, the probability of answering the item correctly for masters of the node should be significantly higher than for non-masters of the node. As shown in *Table 3*, Item *21793* is flagged because the main effect value is smaller than .5, and thus the item is non-informative from a statistical sense. The probability of answering Item

21793 correctly with mastery of the node ($prob_{mastery}$ = .76) is not significantly higher than that without mastery of the node ($prob_{nonmastery}$ = .68).

Table 3. Example of conditional item probability with predicting node ELA-1141

Items	Intercept	Main effect	Non-Master		Master	
			Correct	Incorrect	Correct	Incorrect
14208	-1.25	1.18	0.35	0.65	0.63	0.37
14210	-0.24	1.34	0.59	0.41	0.85	0.15
14212	-1.12	2.86	0.38	0.62	0.91	0.09
14497	-2.79	3.35	0.18	0.82	0.86	0.14
14498	-6.94	7.61	0.00	1.00	0.88	0.12
14499	-13.97	15.08	0.00	1.00	0.92	0.08
21790	0.26	1.89	0.67	0.33	0.93	0.07
21791	0.83	142.34	0.79	0.21	1.00	0.00
21793	0.30	0.38	0.68	0.32	0.76	0.24
21815	-0.75	1.80	0.46	0.54	0.84	0.16
21816	0.46	264.03	0.74	0.26	1.00	0.00
21817	-0.24	24.50	0.58	0.42	1.00	0.00

After items are flagged for review, the next step is for content experts to review the content of the items to determine what changes, if any, are needed to better assess the node. Decisions concerning the structure of the learning map should never be made on statistical evidence alone. Rather, statistical findings should be used as one tool for evaluating the structure of the learning map, taken into consideration with theoretical research regarding skill acquisition in the areas of English language arts, as well as expert judgments from content specialists on the order and size of learning targets.

Once statistical evidence concerning recommendations for revisions to the map structure was compiled, content teams examined the evidence to make final judgments regarding revisions to the nodes and their connections, with statistical recommendations serving as just one piece of evidence in their final decision. Final content decisions about map structure were made based on converging evidence from multiple sources. Because the nodes and connections in the learning

map represent knowledge, skills and abilities that span from simple cognitive behaviors to complex evaluative tasks, statistical evidence provides one source of information for evaluating the structure of the map; however decisions must be considered in the context of the research that informed the original ordering of nodes and connections.

In English language arts, one content-related consideration used to evaluate recommendations about map structure was the distinction between cognitive processes that underlie comprehension and the products of comprehension. Since the focus of dynamic assessments of emergent and conventional literacy are on the cognition that underlies comprehension, much of the research base that informed the original structure of the learning map was based on empirical studies that described both processes and products of comprehension as well as the cognitive behaviors that lead to emergent and conventional literacy. Zwaan and Singer (2003, p. 85) describe "online" methods that are used in text comprehension research to measure cognitive processes during reading rather than afterward. These "activation measures" are used to measure the availability of information to a reader as he or she comprehends a text. Since there are distinctions in the learning map between online processes and offline products of comprehension it is important to evaluate statistical information in light of the original intent of the node; does it represent a comprehension process or a product of comprehension?

An additional consideration is the context of the assessment. It is difficult to measure a student's ability to comprehend text without giving her a text to read. Traditional measures of reading comprehension rely on students reading a text, mentally representing the information in the text, and finally recalling relevant portions of the mental representation or rereading the text to find relevant information in response to questions. The assessment system uses multiple types

of passages to support assessment. Potential sources of item difficulty can, in some cases, be ascribed to variation in passage quality and complexity, which should be accounted for before making alterations to the structure of the learning map.

Significance

The use of a learning map for dynamic assessment is based on the premise that the map itself, consisting of skills and pathways, is correct. Any errors in the specification of the map can distort the assessment of skill development and lead to misdiagnosis of student learning states. The research presented here functions as one source of validation the English language arts portion of the leaning map, The findings of this research provide statistical evidence of the correct specification of the learning map, and support the inferences to be made from student performance based on items measuring the nodes.

Findings from studies such as this can also be used to support test development. Test developers are able to use the modeling results to inform the any future revisions of the learning map in order to ensure the best possible model of skill acquisition. Similarly, statistical evidence obtained from diagnostic modeling can also be used to make decisions about how content should be assessed as part of the assessment system. Test developers can evaluate which items and nodes are working well in the context of the map and which items and nodes tend to result in flags, thereby informing future development efforts.

One limitation of the current research is that it only examined a portion of the English language arts part of the learning map. A similar study is needed in mathematics in order to support inferences being made regarding student skill acquisition. In addition, modeling

considerations should be given to areas of the map that may have a smaller sample size than the threshold of 100 used in this study.

The research presented here expands on previous scholarly work in the area of diagnostic assessment. An algorithm of this scope and magnitude has not previously been documented in the literature. Research on diagnostic assessment typically includes a substantially smaller number of nodes and connections, with thirty to sixty nodes or attributes being at the high end. Because of the size and scale of the cognitive skills reflected in this dynamic assessment system, findings are likely to impact future modeling efforts and lead to further benefits to students and teachers in classrooms.

References

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule-space model and attribute hierarchy method. *Journal of Educational Measurement, 44*, 325-340. doi: 10.1111/j.1745-3984.2007.00042.x

Heidelberger, P., & Welch, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM, 24*(4), 233-245.

Heidelberger, P., & Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research, 31*(6), 1109-1144.

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

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331). New York: American Council on Education/Macmillan.

Trout, D. L., & Hyde, B. (2006, April). *Developing score reports for statewide assessments that are valued and used: Feedback from K-12 stakeholders*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Yang, X., & Embretson, S. E. (2007). Construct validity and cognitive diagnostic assessment. In J. P. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education* (pp. 119-145). New York: Cambridge University Press.

Zwaan, R.A. & Singer, M. (2003). Text Comprehension. In A.C. Graesser, M.A. Gernsbacher & S.R. Goldman (Eds.), *Handbook of Discourse Processes.* Mahwah, NJ: Lawrence Earlbaum.