Symposium on

Beyond Learning Progressions: Maps as Assessment Architecture

Title:

Empirical Methods for Evaluating Maps: Illustrations and Results

W. Jake Thompson and Brooke Nash

University of Kansas

Author Note

Paper presented at the 2019 annual meeting of the National Council of Measurement in Education, Toronto, Canada. Correspondence concerning this paper should be addressed to W. Jake Thompson, ATLAS, University of Kansas, 1122 West Campus Road, Lawrence, KS, 66045; jakethompson@ku.edu. Do not redistribute this paper without permission of the authors.

Abstract

Learning progressions and learning map structures are increasingly being used as the basis for the design of large-scale assessments. Of critical importance to these designs is the validity of the map structure used to build the assessments. Most commonly, evidence for the validity of a map structure comes from procedural evidence gathered during the learning map creation process (e.g., research literature, external reviews, etc.). However, it is also important to provide support for the validity of the map structure with empirical evidence using data gathered from the assessment. In this paper, we propose a framework for the empirical validation of learning maps and progressions using diagnostic classification models. Three methods are proposed within this framework that provide different levels of model assumptions and types of inferences. The framework is then applied to the Dynamic Learning Maps® (DLM®) alternate assessment system to illustrate the utility and limitations of each method. Results show that each of the proposed methods have some limitations, but are able to provide complementary information for the evaluation of the proposed structure of content standards (Essential Elements) in the DLM assessment.

Keywords: learning maps, learning progressions, diagnostic classification models

Empirical Methods for Evaluating Maps: Illustrations and Results

Learning progressions (LPs; also known as learning trajectories [LTs]) are a model of pedagogical thinking to describe shifts between understanding new knowledge and more advanced targets as a sequence of possible transformations (Simon, 1995). LPs grew out of the Science Education for Public Understanding Program (SEPUP; Roberts, Wilson, & Draney, 1997) with "construct maps", which are conceived of as "strategically developed cycles and sequences of instructional activities that guide learning pathways" (Duschl, Maeng, & Sezen, 2011, p. 131). Thus, LPs are meant to describe the process of acquiring new knowledge over a given period of time, or within a specific learning or content area (National Research Council [NRC], 2007).

LPs in science and mathematics education are currently seen as promising strategies for the redesign and reform of curriculum, instruction, and assessment in educational environments (Corcoran, Mosher, & Rogat, 2009; Duschl et al., 2011). LPs mainly rely upon cognitive science research on how students learn a particular concept to describe a path of skill acquisition (Alonzo & Steedle, 2009). The NRC volume *Knowing What Students Know* (NRC, 2001) recommends the use of cognitive models take a central role in the assessment design process. Consistent with this idea, LPs can provide a framework for the development of both large-scale and classroom-based assessments to measure how understanding develops in a given domain.

Validating the Structure of LPs

Of critical importance when considering the potential uses of LPs in practice is a validation of the proposed structure. If the proposed structure is incorrect, then any inferences and instructional decisions that are made as a result of the proposed structured are at risk of being incorrect as well. Traditionally, evidence supporting the structure of LPs has fallen into two categories. Procedural evidence is focused on the process of how the proposed structure of the LP was created. Empirical evidence is focused on statistical methods that can be used to validate the LP structure once data has been gathered to measure the proposed knowledge/skills and their connections. Both types of evidence are described below, although the main focus of the remainder of the paper is empirical evidence.

Procedural Evidence Approaches. Procedural evidence refers to the process of creating the LP(s) prior to the collection of any data. This involves the initial research and literature review that goes in to defining the skills and how they are connected, external review of the LPs by individuals not directly involved in the initial creation, and alignment studies. This has been the approach taken by many projects. For example, SEPUP (Roberts et al., 1997) created LPs with educator input, scoring guides, and examples of student works along the progression. By following the increasing complexity of the associated materials, it is possible to also follow the logic of the LP organization. Similarly, the Teacher Analysis of Student Knowledge (TASK; Supovitz, Ebby, & Sirinides, 2013) developed a progression for teachers' understanding of their students' mathematics knowledge. This progression was then tied back to the Common Core State Standards, providing procedural evidence through the vertical articulation of the standards and the development of the LP. The Dynamic Learning Maps® (DLM®) alternate assessment project followed similar approaches when developing the learning map models for English language arts, mathematics and science (Andersen & Swinburne Romine, 2019; Swinburne Romine & Schuster, 2019).

Procedural evidence is undoubtedly important to any LP that is developed. Without theoretical knowledge to support the proposed structure that makes sense conceptually, empirical evidence is unlikely to be sufficient. Indeed, having an entirely data driven LP may result in a structure that is overfit to the collected data or conflicting with the wider research literature. However, procedural evidence is also insufficient in isolation. Although LPs can be developed using all best practices, there is always some level of uncertainty in the development of the structure. Thus, it is important to collect data and provide empirical evidence to corroborate the proposed structure.

Empirical Evidence Approaches. Empirical evidence involves collecting and analyzing data to evaluate the structure of an LP. There are many forms that this type of evidence could take. One example is analyzing item responses that align to specific levels within an LP. This was the approach taken by Briggs, Alonzo, Schwab, & Wilson (2006), who developed an assessment using ordered multiple-choice items to assess students' level of achievement within a learning progression. In this assessment design, answer options are tied to specific levels in the progression, and the aggregated response patterns can provide evidence to support the structure (i.e., across all levels of assessment, a student's selected responses should consistently correspond to a level in the progression).

Another possibility for empirical evidence is to relate the proposed LP to external outcomes. In the Length Measuring Learning Trajectory (LMLT), Barrett, et al. (2012) showed that the use of their LPs in instruction were able to predict student growth within the targeted skills, and that the use of the LMLT in formative instruction was associated with higher achievement in those areas during a final assessment. A similar approach was used by Jin, Shin, Johnson, Kim, & Anderson (2015) to develop LPs of science content. Using item response theory (IRT), Jin et al. (2015) first calibrated separate models for each level of the LP to show the distinctness of the progression. They then also showed that as teachers' understanding of the LP increased, so did the performance of their students on the post-test. Thus, as teachers better instruct their students along this progression, students demonstrate better understanding of the material. This approach of using external data for evidence was also employed by Supovitz, Ebby, Remillar, & Nathenson (2018) on the Teacher Analysis of Student Knowledge (TASK). In this project, teachers were asked to blindly rate student responses to items measuring different levels of a mathematics LP. Teachers were then asked to place their students on the LP in the location that best represented their acquired knowledge. The student responses were then compared to the teacher placement in the LP. The findings showed that teachers were able to place students on the LP consistent with the students' item responses.

Empirical evidence can also take the form of classical item statistics (i.e., p-values). Herrmann-Abell & DeBoer (2018) developed an LP for the complexity of energy in science. Items were written at three levels of complexity for each stage of the LP. They then used an ANOVA to compare the mean item p-values for each level across the different stages. The results indicated that the proposed order of the stages was not supported by the data. However, Kendall's τ correlations indicated that second progression, one of conceptual complexity that did follow the proposed stages. Thus, the totality of the evidence suggests that the stages of the LP are increasing complex, but not in a way that necessitate acquisition of one before another.

A similar approach was taken by Clark, Kingston, Templin, & Pardos (2014) in an early evaluation of the DLM pilot administration. In this study, students were grouped into four complexity bands (Foundational, Band 1, Band 2, and Band 3) based on the students' expressive communication behaviors and subject knowledge, as reported by their teachers. In the pilot study, students were administered test items at multiple levels, corresponding to different areas of the learning map structure. Within a complexity band, Clark et al. (2014) observed that the probability of providing a correct response decreased as the level of the testlet increased, thus providing preliminary evidence of the general ordering of the structure.

Taken together, empirical evidence has the ability to not only support (or refute) the proposed LP structure, but also provide unexpected findings and avenues for further research. However, these methods are all somewhat limited, in that none use a model-based method that is consistent with the multidimensional nature of LPs. For example, when using unidimensional and multidimensional IRT models, the different stages of the LP can be modeled as separate latent abilities, but the structure of the LP cannot be fully enforced (see Deng, Roussos, & LaFond, 2017; Schwartz, Ayers, & Wilson, 2017). Additionally, although the methods are useful for linear LPs, they may not generalize to a more complicated learning map structure, where multiple pathways can lead to the same knowledge acquisition. Thus, a flexible and generalizable framework of empirical validation is needed to fully evaluate these structures.

Map Validation with Diagnostic Classification Models

The purpose of this paper is to describe and illustrate with examples, empirical approaches to learning map (and LP) validation using diagnostic classification models (DCMs). Specifically, three methods with varying levels of complexity and model assumptions are defined and then illustrated using the DLM alternate assessment. The methods are described in the context of the DLM learning maps; however, these methods generalize to other learning map or LP models as well.

Diagnostic Classification Models

DCMs (also known as cognitive diagnostic models), are a class of multi-dimensional psychometric models that define a mastery profile on a predefined set of attributes (Rupp & Templin, 2008; Rupp, Templin, & Henson, 2010). Given an attribute profile for an individual, the probability of providing a correct response to an item is determined by the attributes that are required by the item. Whereas traditional psychometric models (e.g., IRT) model a single, continuous latent variable, DCMs model student mastery on multiple latent variables or skills of interest. Thus, a benefit of using DCMs for calibrating and scoring operational assessments is their ability to support instruction by providing fine-grained reporting at the skill level. Based on the collected item response data, the model determines the overall probability of students being classified into each latent class for each skill.

DCMs can also be used to test different learning map or LP structures. Given a number of attributes (e.g., nodes in a map, or stages/steps in an LP), there are 2^A possible attribute profiles, where *A* is the number of attributes. This represents all possible combinations of mastery and non-mastery across the attributes. By limiting the number of possible profiles, it is possible to test different structures. For example, Templin & Bradshaw (2014) used a hierarchical DCM to test an attribute structure of skills related to English grammar, where some attributes had to be mastered in order for other attributes to be mastered. This hierarchical model adapts the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2008), which is discussed in more detail.

The Log-linear Cognitive Diagnostic Model. The LCDM provides a generalized DCM that subsumes many of the other more restrictive DCMs. For example, the deterministic-input noisy-and-gate (DINA; de la Torre & Douglas, 2004) and deterministic-input noisy-or-gate (DINO; Templin & Henson, 2006), are both subsumed by the LCDM (Henson et al., 2008). In the LCDM, the probability, π, of an individual in class *c* providing a correct response to item *i* can be expressed as:

$$\text{logit}(\pi_{ic}) = \text{logit}(P(X_{ic} = 1|\boldsymbol{\alpha}_c) = \lambda_{i,0} + \sum_{a=1}^{A} \lambda_{i,1,(a)}\alpha_{ca}q_{ia} + \sum_{a=1}^{A}\sum_{a'>1} \lambda_{i,2,(a,a')}\alpha_{ca}\alpha_{ca'}q_{ia}q_{ia'} + \cdots$$

where α_{ca} is a binary indicator for whether or not attribute a is mastered for respondents in class c, and q_{ia} is a binary indicator for whether or not attribute a is measured by item i. For the estimated parameters, $\lambda_{i,0}$ represents the intercept, $\lambda_{i,1,(a)}$ is the simple main effect for attribute a on item i, and $\lambda_{i,2,(a,a')}$ is the two-way interaction between attribute a and a' on item i. Further interactions can be added as necessary for items that measure more than two attributes.

The specification of a DCM has several advantages. First, the parameters are straightforward to interpret, as they are on the same log-odds scale as a standard logistic regression. Additionally, the LCDM parameters can be restricted in order to estimate other models for model comparison. For example, if all parameters except the intercept and highest-order interaction for each item are fixed at 0, the model is equivalent to the DINA model. Thus, multiple models with different assumptions can be fit using the same framework and directly compared. Finally, as mentioned previously, the LCDM can be extended to test attribute hierarchies. Thus, this is the framework used for map validation for the DLM alternate assessment.

The Dynamic Learning Maps Alternate Assessment

The DLM assessments are built based on learning map models, which are a type of cognitive model consisting of interconnected learning targets and other critical knowledge and skills (DLM Consortium, 2016). The development of the learning map models started with a literature review, including existing learning progressions that typically describe a single, linear sequence of building block skills towards an academic grade-level target skill for general education students. Using published research, the map was expanded to include alternate and additional pathways intended to provide students with significant cognitive disabilities with multiple routes to achieving learning targets. That is, rather than conceptualizing knowledge acquisition in a linear path (as in traditional LPs), the DLM learning maps consist of multiple pathways that can be followed in order to realize higher skills (DLM Consortium, 2016).

In the DLM assessment, Essential Elements (EEs) are specific statements of knowledge and skills, and are the learning targets for the assessment. For each EE, there is an associated mini-map of nodes from the full learning map. These nodes represent critical junctures on the path to, at, and beyond the standard identified in the EE. Within these maps, neighborhoods of related nodes are grouped together into linkage levels, which is the level of analysis for scoring and reporting in the DLM operational assessment. Although multiple nodes may be grouped into the same linkage level, a single node will never belong to multiple linkage levels within the same EE. For ELA and mathematics, there are five linkage levels within each EE: Initial Precursor, Distal Precursor, Proximal Precursor, Target, and Successor. In science, there are only three linkage levels within each EE: Initial, Precursor, and Target. These linkage levels are assumed to follow a hierarchical structure, whereby higher linkage levels can only be mastered if the lower levels have also been mastered. Conversely, if a lower linkage level has not been mastered, it is impossible to master higher linkage levels. Finally, items are written to specific nodes in the mini-map, and administered together on a testlet with other items that measure the same linkage level. Because a linkage level may contain more than one node, multiple nodes may be measured on a single testlet. However, a testlet will only measure one linkage level. The relationship of nodes, linkage levels, items, and testlets within an EE can be seen in Figure 1 (reproduced from DLM Consortium, 2016).

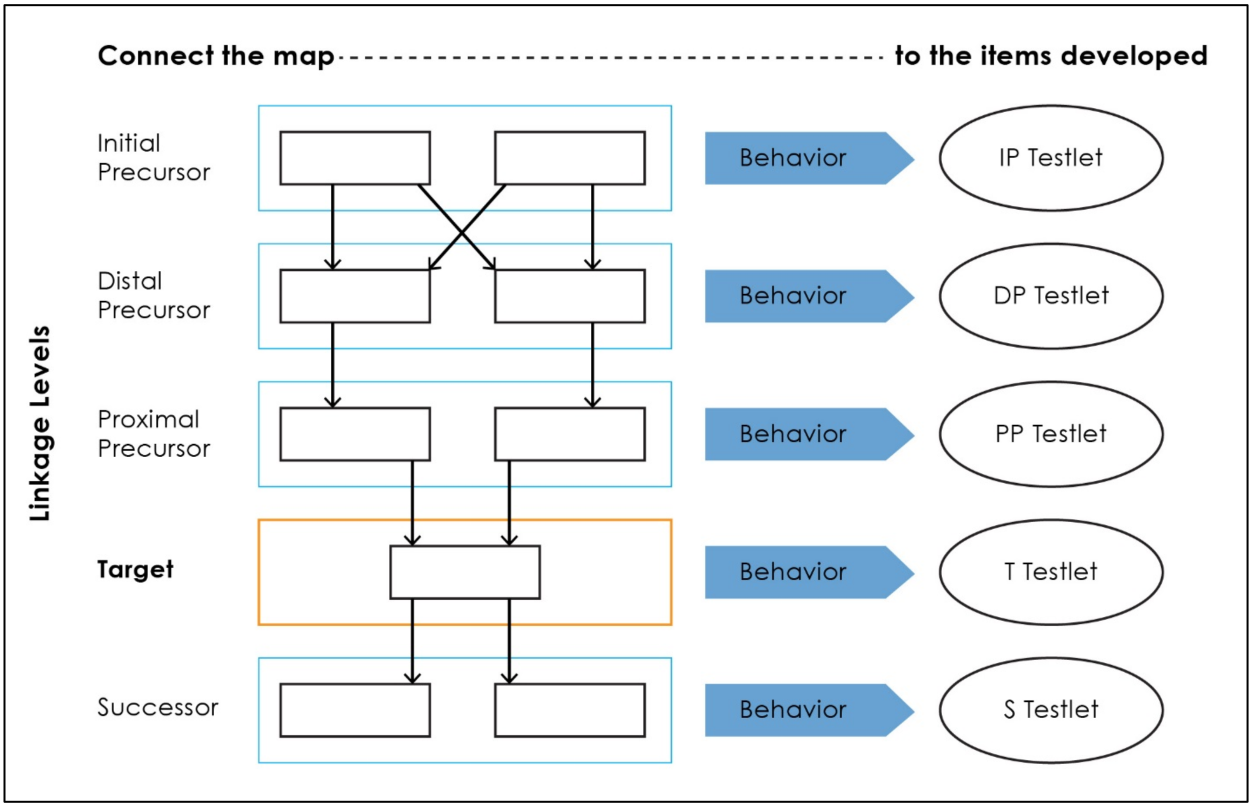

Figure 1. Relationship between DLM map nodes in five linkage levels and items in testlets. Small black boxes represent nodes in the DLM map. Blue and orange boxes represent collections of nodes in linkage levels. The orange box denotes the Target linkage level for the EE.

In the spring operational assessment, students are administered a series of testlets, with each testlet measuring one linkage level from one EE. The level of the first testlet a student tests on is determined by the system using information provided by the student's teacher. After this initialization assignment, subsequent levels are assigned based on student performance on the previous testlet. For example, in Figure 2, the student started their first testlet at the Distal Precursor level. Following the first testlet they adapted up to the Proximal Precursor level in testlet two, and adapted up again to the Target level for testlet three. They then adapted back down to the Proximal Precursor level in testlet four, where they stayed for the final testlet as well. Critically, this adaptation *does not* happen within EE, but rather across testlets. EEs that are tested in one testlet are not repeated in other testlets within the operational assessment. Thus, this student would test on EE #1 at the Distal Precursor level, EE #2 at the Proximal Precursor level, etc. The result of this assignment process is that students end up testing on only one linkage level within each EE.

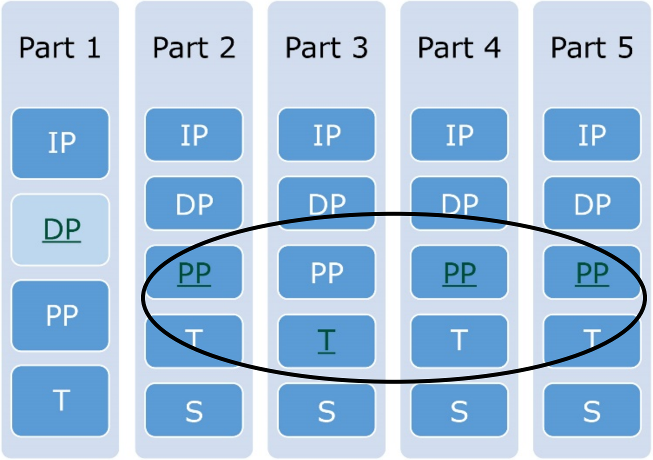

Figure 2. Linkage levels adapting up and down between testlets.

In summary, the DLM data is collected such that each item measures one linkage level. Items are nested within testlets, and all items on a testlet measure the same linkage level. Finally, adaptation occurs in the spring assessment between testlets, and therefore between EEs, not within EEs. Thus, there is no built-in mechanism on the operational assessment for the collection of data on multiple linkage levels within an EE for a single student. Instead, this data is collected through other means, such as field tests that intentionally assess students at a different linkage level than what was assessed during the operational assessment.

To date, the structure of the DLM map model has been validated primarily with procedural evidence (see Andersen & Swinburne Romine, 2019; Swinburne Romine & Schuster, 2019). Empirical approaches were initially limited due to data sparseness inherent in the operational design (i.e., no vertical data within an EE). However, recent data collection efforts have focused on collecting data for students at multiple linkage levels. Due to this increase in data across linkage levels, empirical studies can be conducted to evaluate the hierarchical assumptions of the linkage level ordering.

Methods

Under the DCM framework of map validation, three methods are defined for testing a map structure: patterns of mastery profiles, patterns of attribute mastery, and patterns of attribute difficulty. All methods were then applied to the DLM assessments for ELA, mathematics, and science. To demonstrate the methods, analyses are focused on the assumption of the linear hierarchy of linkage levels with EEs.

Method 1: Patterns of Mastery Profiles

Method 1 is the most model based of the three methods. This method involves estimating two competing DCMs. The first model contains all possible profiles, and the second model with only the profiles hypothesized by the map structure. Take for example linear structure of three attributes (such as in the DLM science assessment). With three attributes, there are $2^3 = 8$ possible mastery profiles. However, under the hierarchical assumption of the linkage levels, not all of those profiles are possible. If the hierarchical structure is correct, only the monotonically increasing profiles should be possible. This is illustrated in Table 1.

Table 1

Possible and Hypothesized Mastery Profiles

Profile	**Initial**	**Precursor**	**Target**
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1
5	1	1	0
6	1	0	1
7	0	1	1
8	1	1	1

Note. Monotonically increasing hypothesized profiles are shaded.

A Bayesian estimation is used to estimate the two LCDM models. The saturated model contains all possible mastery profiles. In the reduced model, only the hypothesized profiles are allowed. Estimation is performed using the RStan interface (Stan Development Team, 2018) to the *Stan* probabilistic programming language (Carpenter, Gelman, Hoffman, Lee, Goodrich, Betancourt, Brubaker, Guo, Li, & Riddell, 2017). If the hierarchical structure holds, then it would be expected that the two models have comparable model fit. If the reduced model fits significantly worse than the saturated model, then this is evidence that the proposed structure is too restrictive. Absolute model fit is assessed through posterior predictive model checks (PPMC). Relative fit is assessed through cross validation.

Absolute fit. Absolute measures of model fit measure the extent to which the model actually fits the data. In the proposed method, absolute fit is assessed using PPMC. PPMC involves generating simulated replicated data sets, creating summary statistics for each replicated data set, and then comparing the distributions of the summary statistics to the value of each statistic in the observed data (for more details and examples see Gelman, Carlin, Stern, Dunson, Vehtari, & Rubin, 2014; Levy & Mislevy, 2016; and McElreath, 2016). In a Bayesian estimation process, a posterior distribution is generated for each parameter in the model. The size of the posterior sample depends on the length of the Markov-Chain Monte Carlo (MCMC) chains. In these models, four chains are estimated with 2,000 iterations each, and the first 1,000 are discarded for the warm-up period. This results in 4,000 retained draws (1,000 from each chain) that make up the posterior samples for each parameter. Thus, 4,000 replicated data sets can be created. Each data set is generated using the values of the parameters at a given iteration. This means that the uncertainty in the parameter estimates is incorporated into the simulation of the replicated data sets. Further, these replicated data sets represent what the data would be expected to look like *if the estimated model is correct*. Thus, deviations in the observed data from the replicated data sets would indicate model misfit. Once the replicated data sets have been generated, summary statistics can be calculated.

Two main statistics are used for assessing the fit of the model. The first is item p-values. For each replicated data set, the p-value for each item is calculated. This results in a distribution of the expected p-value for each item. The p-values from the observed data are then compared to the data. If the observed p-value for an item falls outside of the middle 95% of the expected distribution (or any other interval that is desired; referred to as a credible interval), then the item is flagged for model fit. To summarize the model as whole, the number or percentage of flagged items can be calculated.

In addition to item p-values, an expected distribution of raw scores can be calculated. For this summary, the number of students at each raw score point (sum score across items) is calculated, resulting in a distribution of the expected number of students at each score point. The observed number of students at each score point is then compared to these distributions using a 95% credible interval. For a global evaluation of model fit, this summary can be taken one step further. The mean of the distributions for each score point can be thought of as the expected number of students for that score point. These expected counts can be used to calculate a χ^2-like goodness-of-fit statistic. This is calculated in the same way as a traditional χ^2 statistic, but will not follow the distributional assumptions of the χ^2. However, an empirical distribution can be estimated using the replicated data sets. For each replicated data set, the χ^2 is calculated using the number of students observed at each score in that replication and the expected counts (the means of the distributions). Thus, a χ^2 is estimated for each replication, creating a distribution of expected χ^2 statistics. The χ^2 from the observed data is then compared to this distribution, and a posterior predictive p-value (*ppp*) is calculated as the proportion of the empirical χ^2 distribution that is greater than the observed χ^2. If the *ppp* is less than .05 (or another chosen threshold), then the model is rejected (i.e., the observed data is inconsistent with the replicated data sets, and therefore the model fit is not sufficient).

Absolute model fit is crucial to the evaluation of any model, including DCMs. If the model does not fit the data sufficiently, then any inferences made from the model are prone to error. However, absolute fit indices are unable to adequately compare competing models. For example, if two models both show sufficient absolute fit, these indices are unable to differentiate which should be preferred. For this analysis, relative fit indices are needed.

Relative fit. Relative model fit indices directly compare two models to determine which provides a better overall fit to the data. These are common measures in many models outside of DCMs. For example, the Akaike information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1978) are widely used and recognized. Another relative fit comparison is cross validation. In cross validation, a proportion of the data is withheld from the estimation process, and fit assessed on the held-out portion. This is then repeated multiple times with different training and testing sets. The performance of the model across the held-out portions is then compared between models. In the proposed method, an approximation of leave-one-out cross validation known as Pareto-smoothed importance sampling leave-one-out cross validation (PSIS-LOO; Vehtari, Gelman, & Gabry, 2017a; Vehtari, Gelman, & Gabry, 2017b) is used. This method estimates the predictive density of the model, balancing predictive power with model complexity. This method is also readily available for models estimated with RStan using the loo package (Vehtari, Gabry, Yao, & Gelman, 2018). The loo package also provides the widely applicable information criterion (WAIC; Watanabe, 2010). The PSIS-LOO and WAIC are asymptotically equivalent; however, the PSIS-LOO is more robust when non-informative priors are used, or when there are influential observations.

When examining the PSIS-LOO, the magnitude of the difference in the expected log predictive density (ELPD; the predictive power) compared to the standard error of the difference. If the magnitude of the difference is much larger than the standard error (e.g., 2x as large), then one model is preferred over the other, with the sign of the difference indicating the preferred model. If the difference is negative, the first model in the comparison is preferred. A positive difference indicates the second model in the comparison. Thus, this test is symmetric. If a comparison of the saturated to the reduced has a difference in the ELPD of 20, then a comparison of reduced to saturated model will have an ELPD difference of -20.

Although relative fit indices can provide information about which model may be preferred in a comparison, these values are not useful in isolation. An ELPD from the PSIS-LOO is dependent on the size of the sample and the likelihood function, and therefore not comparable across different types of models or data sets. Additionally, these methods do not tell you if the model fits the data. The comparisons are all relative to the other models. For example, the PSIS-LOO may indicate a preference for the reduced model over the saturated model; it could be that both models fit poorly, but the reduced model is less poor. Therefore, it is important that these methods be used in conjunction with absolute fit indices in order to ensure a comprehensive assessment of model fit.

Method 2: Patterns of Attribute Mastery

Method 2 removes some of the model complexities that exist in Method 1. Rather than estimating the LDCM with mastery profiles across all attributes, a separate LCDM is estimated for each attribute. This is similar to the approach taken by Jin et al. (2015). Using this method, each attribute is treated as independent of the other attributes. Thus, if there are five attributes or nodes in the learning map, five LCDM models would be estimated. Each LCDM then has two possible classes: master and non-master of the given skill. With only two classes, this model is equivalent to a latent class analysis (see Bartholomew, Knott, & Moustaki, 2011). These models are again estimated using RStan (Stan Development Team, 2018). After the estimation of the models, the probability of each student being a master of each skill is calculated. Thus, each student will have a probability of mastery for each attribute that they tested on, calculated from the separate model estimations.

Patterns are then examined across the attribute masteries. The patterns of the probabilities can be compared directly, or dichotomized into 0/1 mastery decisions using a mastery threshold (e.g., .8). If the hierarchical structure holds, then it would be expected the probability of mastery would decrease as the learning map progressed to higher levels. Similarly, if using a dichotomized score, a student should not receive a “master” classification on an attribute unless all the lower levels also received a “master” classification.

To assess the map structure overall, the percentage of students that have an unexpected pattern of attribute mastery can be calculated. The expectation is that the percentage would be low if the hierarchical structure is correct. Further analyses can also examine where the structure may be incorrect. For example, if students are commonly getting flagged for an incorrect pattern of attribute mastery between the second and third attributes, this may be a specific focus of further research and investigation. Flagging can be done at the EE level, or at specific junctions of levels. For example, in these analyses, a structure may be flagged if more than 25% of students have an unexpected pattern within an EE (flagging for overall EE). Alternatively, a structure may be flagged if more than 50% of students had an unexpected pattern across an adjacent pairing of linkage levels (flagging for specific linkage level ordering within EE). For this study, the lower 25% threshold for flagging an EE was chosen to be more conservative for flagging. As the current study is exploratory, the decision was made to error on the side of over-reviewing flags, rather than under-reviewing. Because the sample of students who tested on a given pair of linkage levels is necessarily smaller than then the full sample of students who tested on an EE, more noise may be expected in these comparisons. Thus, a higher threshold was chosen for this criterion to avoid unnecessary flagging due to noisy data.

Method 3: Patterns of Attribute Difficulty

Method 3 represents another step down on the scale of model dependency. Whereas Method 1 and Method 2 both use some version of the LCDM, Method 3 does not depend on the estimation of a model. Similar to Herrmann-Abell & DeBoer (2018), Method 3 involves the calculation of item difficulties for each node or attribute, and then comparing the pattern of difficulties across attributes within student cohorts. Within each cohort, it is expected that the items should get harder as the level increases.

For each item within each cohort, the p-value for the item, *i*, is defined as:

$$p_i = \frac{1}{n_i}\sum_{i=1}^{n_i} X_i$$

Where n_i is the number of students in the cohort who tested on item *i,* and X_i is the score on item i. The standard error of the item p-value for the cohort is then defined as:

$$se_i = \sqrt{\frac{p_i(1-p_i)}{n_i}}$$

A weighted average p-value is then calculated for all items with a level for the cohort. The weight for each item, w_i, is given as:

$$w_i = \frac{1}{se_i^2}$$

The weights for all items within a level for the cohort are then scaled to sum to 1.0. The weighted average p-value, $\bar{p}$, and weighted standard error, $se_{\bar{p}}$, are given as:

$$\bar{p} = \sum_{i=1}^{I} w_i p_i$$

$$se_{\bar{p}} = \sqrt{\bar{p}(1 - \bar{p}) \sum_{i=1}^{I} w_i^2}$$

Thus, for each cohort and level of the learning map, a weighted average p-value and an associated standard error are calculated. These p-values can then be compared within cohorts, with the expectation that the weighted p-values should get lower as the level increases (i.e., items get harder). Cohorts that don’t follow that pattern across levels (beyond the margin of error) are then flagged for a potential violation of the map structure.

Data

To demonstrate this DCM framework of map validation in practice, all three methods were applied to data from DLM assessments from 2015-2016 to 2017-2018. The full sample was filtered to only include testlets that are currently in the operational pool, and only include students that tested on multiple linkage levels within an EE. Data was collected at multiple levels through a field test design in the spring of 2018 that assigned students content at a linkage level adjacent to the level they were assessed in the operational assessment. Additional data constraints were applied for each method.

For Method 1 and Method 2, students were required to have taken at least three items on multiple linkage levels within an EE in order to be included. This was done to support the stability of the estimates of student mastery for the tested linkage levels. Additionally, due to data sparseness issues as a result of the test administration process, only the three EEs with the most cross-linkage level data from each subject for investigation in Method 1. The sample sizes for the nine EEs selected for Method 1 are shown in Table 2. In Table 2, data for ELA.RL.5.1 includes a total of 11,336 students. Of these, 2,776 tested on both the Initial Precursor and Distal Precursor linkage levels, 1,044 tested on only two linkage levels which were non-adjacent, and 226 students tested on more than two linkage levels. The columns may not sum to the total given overlap in the counts (i.e., a student who tested on the Initial Precursor, Distal Precursor, and Proximal Precursor linkage levels would be counted in the IP/DP, DP/PP, and > 2 Levels Tested columns).

Table 2

Summary of Essential Elements Selected for Method 1

	IP/DP	**DP/PP**	**PP/T**	**T/S**	**Two Non-Adjacent**	**> 2 Levels Tested**	**Total N**
ELA.RL.5.1	2,776	3,988	2,974	328	1,044	226	11,336
ELA.RI.7.5	1,034	1,204	699	972	102	278	4,289
ELA.RI.11-12.3	2,354	1,640	1,442	1,917	1,470	394	9,217
M.4.MD.4.b	4,316	3,076	1,206	1,301	973	280	11,152
M.7.G.1	3,524	3,591	2,338	428	2,698	201	12,780
M.8.G.9	6,097	2,096	2,126	1,960	345	245	12,869
SCI.5.PS.3.1	3,614		5,033		3,638	38	8,699
SCI.MS.ESS.3.3	2,679		6,385		2,859	0	9,244
SCI.HS.PS.1.2	4,435		5,024		4,505	1	9,530

Note. IP = Initial Precursor, DP = Distal Precursor, PP = Proximal Precursor, T = Target, S = Successor; For science, IP/DP represents the cross over between Initial and Precursor and PP/T the crossover between Precursor and Target.

For Method 3, students were grouped into cohorts based on their complexity band, which is derived from educator responses to the First Contact survey and determines a student's starting linkage level in the assessment[1]. Further, items were required to have a minimum sample size of 20 students to ensure adequate estimates of item p-values. Starting from the dataset that included all EEs and students who tested on adjacent linkage levels, an iterative filtering process was used to ensure that after filtering items by a minimum sample size, all students still had data on adjacent linkage levels. Specifically, for each EE and complexity band combination, the following repetitive filtering process was conducted until the dataset became stable: (1) remove items with less than 20 students, then (2) remove students with data from only one linkage level. Based on the above selection criteria, data from 138 ELA EEs, 107 mathematics EEs, and 21 science EEs were used for analysis. Table 3 shows the number of EEs that were available for analysis for each subject and complexity band.

Table 3

Number of Essential Elements Analyzed for Each Subject and Complexity Band

Subject	Foundational	Band 1	Band 2	Band 3	Total
English Language Arts	51	118	131	87	138
Mathematics	27	106	107	54	107
Science	3	12	11	17	21

Note. There are 148 total EEs in ELA, 107 EEs in mathematics, and 34 in science.

Results

To demonstrate the DCM framework for learning map and LP validation in practice, the three methods were applied to the DLM assessment data. The results for each method are presented separately.

Method 1: Patterns of Profile Mastery

For each of the nine EEs selected (see Table 2), a saturated model with all possible

[1] For a complete description of the First Contact survey, see Chapter 4 of the *2014–2015 Technical Manual—Integrated Model* (DLM Consortium, 2016) and the First Contact census report (Nash, Clark, & Karvonen, 2015).

mastery profiles and a reduced model with only the monotonically increasing hypothesized profiles were estimated. All models failed to estimate properly. In the majority of cases, this was due to large amounts of missing data. For example, in ELA and mathematics, students most commonly only tested on two of the five linkage levels. Thus, there is around 60% missing data for those students, who have no data on the other three linkage levels within the EE. With this amount of missing data, the models were unable to converge. Future directions to overcome these shortcomings are discussed below.

Method 2: Patterns of Attribute Mastery

For each EE in the DLM assessment (148 ELA, 107 mathematics, and 34 science), a single attribute LCDM was estimated for each linkage level (five model for ELA and mathematics, three for science). Table 4 shows the number and percent of flagged EEs by subject and grade level, using both flagging rules. For example, six 3rd grade ELA EEs were flagged by at least one of the rules, which is 35% of all possible grade three ELA EEs. In total, 59 of the 289 EEs across all grades and subjects were flagged (20%); however, the majority of these EEs came from ELA (*n* = 41; 69%). Of the 59 EEs that were flagged, 2 were flagged for over 25% of students having an unexpected pattern across tested EEs, 46 were flagged for more than 50% of students having an unexpected pattern on a single linkage level combination, and the remaining 11 EEs met both criteria. As an example, 26% of students who tested on SCI.5.PS.2.1 were flagged for an expected pattern, but all linkage level combinations had less than 50% of students flagged. For ELA.RI.3.2, only 14% of students were flagged for an unexpected pattern, but 79% of students who tested on the Target and Successor levels had an unexpected pattern across those two linkage levels.

Table 4

Number and Percent of Essential Elements Flagged for Essential Element or Linkage Level Combination Rule

	Grade/Course									
	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Bio**
ELA	6 (35%)	3 (18%)	4 (21%)	5 (26%)	7 (39%)	10 (50%)	4 (21%)	--	2 (11%)	--
Math	2 (18%)	3 (19%)	2 (13%)	0 (0%)	2 (14%)	1 (7%)	0 (0%)	1 (11%)	1 (11%)	--
Science	--	--	2 (22%)	--	--	2 (22%)	1 (11%)	--	--	1 (14%)

Note. ELA Essential Elements are grade banded for grades 9-10 and 11-12; Science Essential Elements are grade banded for grade 3-5, 6-8, and 9-12.

A summary of the percentage of EEs within each subject that were flagged for specific linkage level combinations can be seen in Figure 3. For example, approximately 60% of all ELA Ees were flagged for reversals at the Target/Successor combination. Put another way, in 60% of all flagged ELA Ees, more than 50% of students who tested on the Target and Successor linkage levels had an unexpected pattern of attribute mastery across those two linkage levels. Figure 3 shows that most flags were due to reversals among the Proximal Precursor, Target, and Successor linkage levels. There were very few Ees flagged due to a reversal between the lowest two levels across all three subjects. This pattern of flagging across linkage levels is not unexpected given the distance between the content being assessed. In general, the content assessed across the Target and Successor levels much closer than the content assessed across the Initial Precursor and Distal Precursor. Therefore, it is more likely to observe a reversal across those linkage levels whose content is closer together.

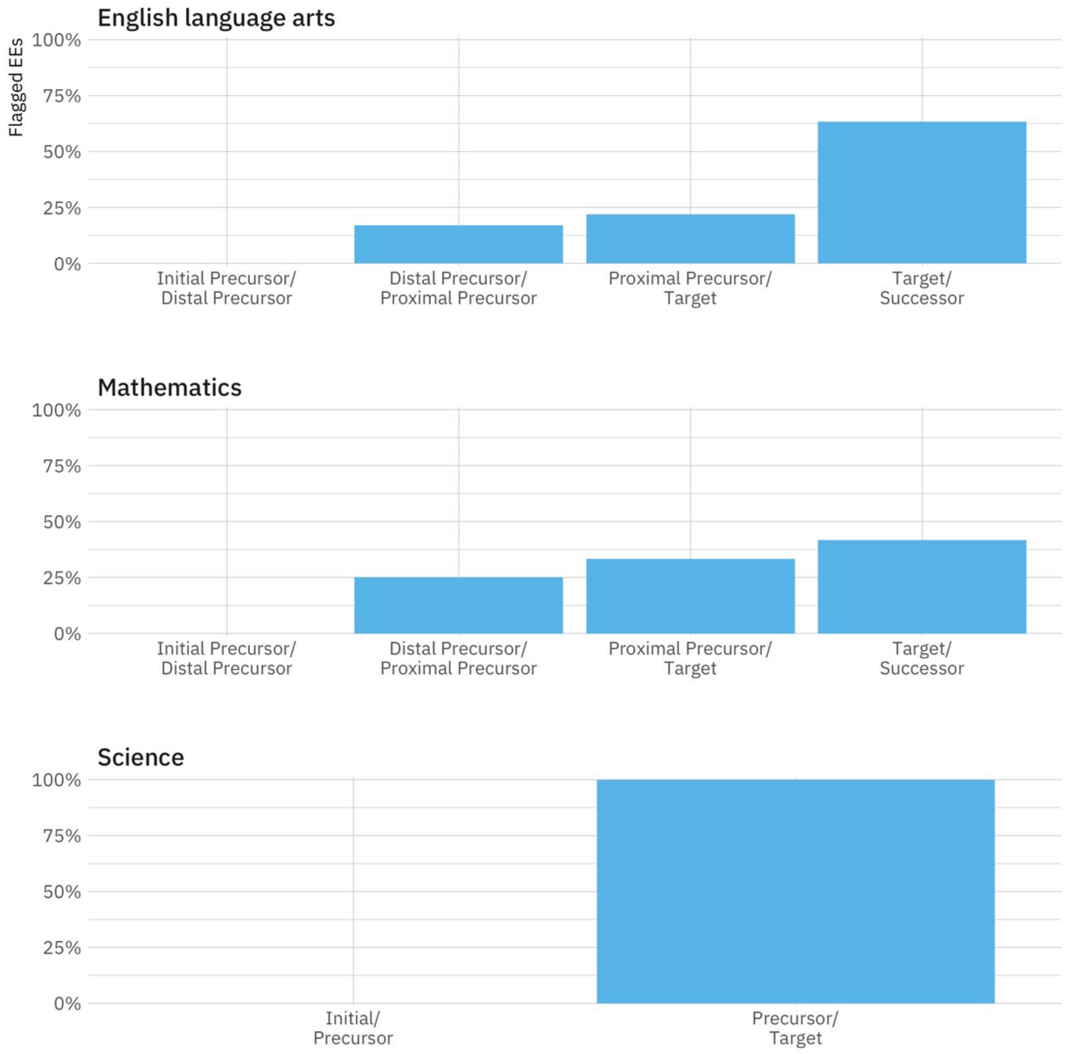

Figure 3. Percentage Essential Elements flagged for reversals at each linkage level combination by subject.

Method 3: Patterns of Attribute Difficulty

Students were grouped into cohorts based on their complexity band, calculated from the First Contact survey, as described in Clark et al. (2014). There are four bands: Foundational, Band 1, Band 2, and Band 3. For each EE, the weighted p-value and standard error were calculated for each cohort. An EE was flagged for a cohort if the weighted p-values did not follow in the expected direction, outside the margin of error, defined as 1 standard error around each weighted average.

Table 5 shows the number and proportion of investigated Ees that were flagged within each subject and complexity band combination across all grade levels. Overall, there are very few inconsistencies. Only 28 total Ees were flagged in ELA and 35 flagged in mathematics. Of the 28 Ees flagged in ELA, 12 (43%) were writing Ees. Three writing Ees were flagged for all four complexity bands, while an additional three writing Ees were flagged in two complexity bands. The remaining 16 ELA Ees were only flagged in one complexity band. Although writing Ees made up a large proportion of flags using when using this method, the same pattern was not seen in Method 2, where only 1 of the 41 flagged ELA Ees was a writing EE.

In mathematics, there was one EE flagged in all four complexity bands, two Ees flagged in three complexity bands, and two Ees flagged in two complexity bands. Notably, no science Ees were flagged for inconsistencies at any complexity band. ELA shows inconsistencies mainly in the Foundational, Band 1, and Band 2 complexity bands, whereas mathematics is mostly confined to Band 1, Band 2, and Band 3.

Table 5

Patterns of Linkage Level Difficulty Flags by Complexity Band and Subject

Subject	**Foundational**	**Band 1**	**Band 2**	**Band 3**
English Language Arts	12 (24%)	13 (11%)	10 (8%)	5 (6%)
Mathematics	2 (7%)	10 (9%)	26 (24%)	6 (11%)
Science	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Note. Sample sizes for each cell are shown in Table 3.

Discussion

In this paper, we present a framework for validating the structure of learning maps and LPs using DCMs. Three methods were described with decreasing levels of model dependency.

These methods were then applied to the hierarchical linear structure of linkage levels within EEs for the DLM alternate assessments for ELA, mathematics, and science to illustrate their use in an operational setting.

In summary, the results from the patterns of attribute mastery (Method 2) and patterns of attribute difficulty (Method 3), using the selected thresholds, demonstrated low flagging rates for the proposed LL structures. Method 2 also indicated potential areas for further evaluation. However, Method 3, though promising and useful, does not offer a model-based assessment of the hierarchical structure. Method 2 also showed relatively low rates of flagging; however, this method also does not directly model the relationships between attributes.

Additionally, the results from Method 2 and Method 3 are dependent on the selected thresholds. In Method 2, the threshold used for determining mastery was the same threshold used in the operational scoring of the DLM assessment (see Chapter 5 of DLM Consortium, 2017), but a different threshold may lead to different results. In addition to the mastery threshold, flagging thresholds in both Method 2 and Method 3 also will also influence the results. In exploratory studies, such as the one described here, lower thresholds may be desired (i.e., higher flagging rates) in order to ensure thorough reviews. For example, a key finding from Method 3 was the high proportion of flags coming from writing EEs. However, these EEs have also been affected by opportunity to learn (DLM Consortium, 2018), and thus there is likely more noise in the data than would normally be expected in an operational program. Therefore, when conducting confirmatory studies used for informing decision making, higher thresholds may be used to ensure flags are warranted.

Finally, Method 1 faced estimation issues due to the large amount of missing data. Thus, despite challenges with applying Method 1, Method 2 and Method 3 do provide preliminary evidence for the hierarchical structure of the linkage levels.

These findings also demonstrate the utility of the proposed framework for map validation using DCMs. By including methods with different levels of model dependency, the less model-dependent methods are still able to provide useful information when the highly model-dependent methods fail to estimate. Additionally, evidence of fit or misfit for the proposed structure from multiple methods provides a more comprehensive set of information from which final inferences can be made. The three methods also provide different levels of detail that may be useful for further examination. For example, when using Method 1, mastery profiles that may be important but ignored in the reduced model can be identified. Using Method 2, specific combinations of attributes that are problematic can be identified. With Method 3, it is possible to evaluate whether certain groups of students are more or less likely to conform to the proposed structure. Thus, these methods provide a flexible framework for evaluating the structure of learning maps and LPs.

Future work will continue to refine the methodology of the DCM framework for map validation and examine the hierarchical structure of the linkage levels from both content-based and model-based approaches. New data collection methods are currently underway to collect more cross-linkage level data through field testing to reduce the amount of missing data present for Method 1. Additionally, modifications to the estimation process in Method 1, such as simplifying the parameterization of the structural model (e.g., Thompson, 2018), are being explored in order to make the method applicable under less than ideal data conditions.

Furthermore, simulation studies are currently underway to evaluate alternative flagging thresholds for individual linkage level pairings and overall EE patterns for Method 2. By simulating data that fits the proposed structure, it is possible to determine what level of flagging would be expected under ideal conditions. Thus, flagging decisions would be based on empirical flagging that may provide a more meaningful criterion for review.

Additionally, Method 3 could be repeated using a different grouping variable. For example, rather than using complexity band, students could be grouped based on previous testlet performance. This may be a better indicator of student equivalence, and thus may provide results that are more actionable from a content perspective. If students were grouped by their most recent testlet performance, the test development team would be able to examine specific testlets from the across the adjacent linkage levels showing possible reversals. Finally, EEs that are flagged across multiple methods will be sent to the test development team to examine the content and underlying mini-maps for possible mis-specifications.

Current and future work is also focused on applying this framework to the finer-grained nodes that make up the linkage levels. Future work in this area includes the DLM assessment system and related projects that are also housed with the Center for Accessible Teacher, Learning, and Assessment Systems (ATLAS; DLM's parent center). The Innovations in Science Map, Assessment, and Report Technologies (I-SMART; Karvonen, 2018) is focused on building node-level science assessments, with the goal of providing feedback to students at the node level. Similarly, the Kansas Learning Map project is a collaboration with the Advanced Innovation Center for Future Education at Beijing Normal University focused on building node-level assessments for proportions and percents in mathematics, and then examining potential structural differences across American and Chinese students. Combined with the studies described here empirically examining the hierarchical assumptions of linkage levels, this area of research has the potential to both inform the literature on student learning processes and provide further guidance for the development of learning map and progression-based assessments.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education*, *93*(3), 389–421. https://doi.org/10.1002/sce.20303

Andersen, L. & Swinburne Romine, R. (2019, April). Iterative design and stakeholder evaluation of learning map models. In M. Karvonen (Moderator), *Beyond Learning Progressions: Maps as Assessment Architecture*. Symposium conducted at the annual meeting of the National Council on Measurement in Education, Toronto, Canada.

Barrett, J.E., Sarama, J., Clements, D., H., Cullen, C., McCool, J., Witkowski-Rumsey, C., & Klanderman, D. (2012) Evaluating and Improving a Learning Trajectory for Linear Measurement in Elementary Grades 2 and 3: A Longitudinal Study. *Mathematical Thinking and Learning 14(1),* 28-54. https://doi.org/10.1080/10986065.2012.625075

Bartholomew, D., Knott, M., & Moustaki, I. (Eds.) (2011). Latent class models. In *Latent Variable Models and Factor Analysis: A Unified Approach* (3rd ed., pp. 157–189). West Sussex, United Kingdom: Wiley.

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

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software 76*(1). doi 10.18637/jss.v076.i01

Clark, A., Kingston, N., Templin, J., & Pardos, Z. (2014). *Summary of results from the fall 2013 pilot administration of the Dynamic Learning Maps™ Alternate Assessment System.*

(Technical Report No. 14-01). University of Kansas, Center for Educational Testing and Evaluation. Lawrence, KS.

Corcoran, T., Mosher, F.A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform.* Consortium for Policy Research in Education Report #RR-63. Philadelphia, PA: Consortium for Policy Research in Education.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353. https://doi.org/10.1007/BF02295640

Deng, N., Roussos, L., & LaFond, L. (2017). Multidimensional modeling of learning progression-based vertical scales. Paper presented at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: a review and analysis. *Studies in Science Education*, *47*(2), 123–182. https://doi.org/10.1080/03057267.2011.604476

Dynamic Learning Maps Consortium. (2016). *2014–2015 Technical Manual—Integrated Model*. University of Kansas, Center for Educational Testing and Evaluation. Lawrence, KS.

Dynamic Learning Maps Consortium. (2017). *2015–2016 Technical Manual Update—Integrated Model*. University of Kansas, Center for Educational Testing and Evaluation. Lawrence, KS.

Dynamic Learning Maps Consortium. (2018). *2017–2018 Technical Manual Update—Integrated Model*. University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems. Lawrence, KS.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (Eds.) (2014). Model checking. In *Bayesian Data Analysis* (3rd ed., pp. 141–164). Boca Raton, FL: CRC Press.

Henson, R. A., Templin, J. L., & Willse, J. T. (2008). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191. https://doi.org/10.1007/s11336-008-9089-5

Herrmann-Abell, C. F., & DeBoer, G. E. (2018). Investigating a learning progression for energy ideas from upper elementary through high school: Learning Progression for Energy Ideas. *Journal of Research in Science Teaching*, *55*(1), 68–93. https://doi.org/10.1002/tea.21411

Jin, H., Shin, H., Johnson, M. E., Kim, J., & Anderson, C. W. (2015). Developing learning progression-based teacher knowledge measures. *Journal of Research in Science Teaching*, *52*(9), 1269–1295. https://doi.org/10.1002/tea.21243

Karvonen, M. (2018). *Innovations in Science Map, Assessment, and Report Technologies (I-SMART)*. Presentation at the 2018 SCILLS & I-SMART Introductory Meeting. https://ismart.works

Levy, R. & Mislevy, R. J. (Eds.) (2016). Model evaluation. In *Bayesian Psychometric Modeling* (pp. 231–252). Boca Raton, FL: CRC Press.

McElreath, R. (2016). Sampling the imaginary. In *Statistical Rethinking: A Bayesian Course With Examples in R and Stan* (pp. 49–70). Boca Raton, FL: CRC Press.

Nash, B., Clark, A., & Karvonen, M. (2015). *First Contact: A census report on the characteristics of students eligible to take alternate assessments*. (Technical Report No. 16-01). University of Kansas, Center for Educational Testing and Evaluation. Lawrence, KS.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: The National Academies Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K–8.* (R.A. Duschl, H.A. Schweingruber, & A.W. Shouse, Eds.). Washington: The National Academies Press.

Roberts, L., Wilson, M., & Draney, K. (1997). The SEPUP assessment system: An overview. BEAR Report Series, SA-97-1. University of California, Berkeley.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research & Perspective*, *6*(4), 219–262. https://doi.org/10.1080/15366360802490866

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications* (1st ed.). New York, NY: Guilford Press. Retrieved from https://www.guilford.com/books/Diagnostic-Measurement/Rupp-Templin-Henson/9781606235270

Schwartz, R., Ayers, E., & Wilson, M. (2017). Mapping a data modeling and statistical reasoning learning progression using unidimensional and multidimensional item response models. *Journal of Applied Measurement, 18*(3), 268-298.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Simon, M. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education, 26,* 114–145.

Stan Development Team. (2018). *RStan: The R interface to Stan*. R package version 2.18.2. https://mc-stan.org/rstan

Supovitz, J. A., Ebby, C. B., Remillard, J., & Nathenson, R. A. (2018). *Experimental Impacts of the Ongoing Assessment Project on Teachers and Students* (Research Report No. RR

2018–1) (p. 16). Philadelphia, PA: Consortium for Policy Research in Education, University of Pennsylvania. Retrieved from https://repository.upenn.edu/cpre_researchreports/107

Supovitz, J., Ebby, C. B., & Sirinides, P. (2013). *Teacher Analysis of Student Knowledge: A Measure of Learning Trajectory-Oriented Formative Assessment* (Synthesis Report) (p. 28). Philadelphia, PA: Consortium for Policy Research in Education, University of Pennsylvania.

Swinburne Romine, R., & Schuster, J. (2019, April). Learning maps as models of the content domain. In M. Karvonen (Moderator), *Beyond Learning Progressions: Maps as Assessment Architecture*. Symposium conducted at the annual meeting of the National Council on Measurement in Education, Toronto, Canada.

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*(2), 317–339.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305. https://doi.org/10.1037/1082-989X.11.3.287

Thompson, W. J. (2018). *Evaluating Model Estimation Processes for Diagnostic Classification Models* (Doctoral Dissertation). Retrieved from ProQuest Dissertations & Theses Global. (Order No. 10785604)

Vehtari, A., Gabry, J., Yao, Y., & Gelman, A. (2018). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.0.0.https://mc-stan.org/loo

Vehtari, A., Gelman, A., & Gabry, J. (2017). Pareto smoothed importance sampling. arXiv preprint arXiv:1507.02646.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413–1432. doi:10.1007/s11222-016-9696-4. arXiv preprint arXiv:1507.04544.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research, 11*, 3571-3594.