DYNAMIC®
LEARNING MAPS

Assessing Model Fit for the Dynamic Learning Maps® Alternate Assessment Using a Bayesian Estimation

Technical Report #18-01

July 2018

Thompson, W. J. (2018). *Assessing Model Fit for the Dynamic Learning Maps® Alternate Assessment Using a Bayesian Estimation* (Technical Report No. 18-01). Lawrence, KS: University of Kansas, Accessible Teaching, Learning, and Assessment Systems (ATLAS).

Contents

Abstract **1**

1 Introduction **2**

2 Description of Models **3**

3 Model Fit Calculation Background **5**
3.1 Absolute Fit 5
3.2 Relative Fit 7

4 Procedure for Evaluating Model Fit **8**
4.1 Data 8
4.2 Method 8

5 Results **8**
5.1 Convergence 8
5.2 Absolute Fit 9
5.3 Relative Fit 11

6 Summary of Model Fit Analyses **11**

References **13**

List of Tables

1 Convergence Rates for All Models 9
2 Average Percent of Items Flagged for Misfit 10
3 Percent of Linkage Levels Flagged for Model Misfit 10

List of Figures

1 Log-odds to probability conversion. 5

2 Example posterior checks for item-level p-values using simulated data. 6

3 Example posterior predictive model check for attribute-level raw score distributions using simualted data. 7

Abstract

As diagnostic classification models become more widely used in operational assessments, it is important to be able effectively evaluate model fit. In this paper, we examine a new method for evaluating model fit using Bayesian model estimation and posterior predictive model checks in the context of the Dynamic Learning Maps® Alternate Assessment System. Our findings suggest that posterior predictive model checks are a methodologically sound way to estimate model fit. However, more work is needed to understand the sensitivity and specificity of these indices. Additionally, more work is needed to evaluate practical significance of model misfit, compared to the statistical significance.

1. Introduction

Assessing model fit is a crucial aspect for any psychometric program. Model fit has important implications for the validity of inferences that can be made from test results. If the model used to calibrate and score the assessment does not fit the data well, results from the assessment may not accurately reflect what students know and can do.

The Dynamic Learning Maps® (DLM®) Alternate Assessment System is based on an interconnected learning map model of discrete skills. The connections between skills indicate the unidirectional ordering of skill acquisition. Nodes in the DLM maps are measured by alternate content standards (Essential Elements; EE), which are of reduced breadth and complexity compared to grade-level college- and career-ready standards. In order to provide all students access to grade-level academic content, each EE is associated with linkage levels which represent the alternate content standard at varying levels of depth, breadth, and complexity. In English language arts (ELA) and mathematics, there are total of five linkage levels for each EE: three precursor linkage levels that lead to the target level and one successor linkage level for students going beyond the target. In science, there are three linkage levels for each EE, two precursor linkage levels and the target level. The availability of multiple skill levels ensures all students are provided access to grade-level content in a way that is most appropriate for the individual student.

One of the key assumptions of the DLM assessment system is that items measuring the same linkage level are fungible, or exchangeable. In this type of model, the item parameters are held constant for all items measuring each linkage level. Thus, the assumption is that a master of the linkage level will have the same probability of providing a correct response to all items measuring that linkage level. Similarly, a non-master would also have the same probability of providing a correct response to all items. Becuase this relationship between the items is assumed by the model, evidence of the degree to which a fungible model fits the data must be evaluated. In addition, the fit of the fungible model should be compared to a non-fungible model to compare their relative fit.

The paper that follows builds on previous work evaluating the model fit of the DLM assessments. An initial investigation using indices based on limited information model fit indices (e.g., Maydeu-Olivares & Joe, 2006; Maydeu-Olivares & Joe, 2014) found insufficient evidence of model fit for the fungible model. However, when examining student score distributions, no significant differences were observed between the fungible and non-fungible scoring models. This suggested that either the model misfit was not practically significant, or the methods were overly sensitive. Due to the sparsity of the data resulting from the DLM administration design (i.e., students typically take testlets that measure only on elinkage level per EE), as well as the potential for over-flagging model misfit, new methods for assessing model fit were developed. Specifically, rather than estimating the model using and expectation-maximization algorithm (Bartholomew, Knott, & Moustaki, 2011), a new Bayesian model estimation was proposed that would allow the evaluation of model fit using posterior predictive model checks (e.g., Gelman et al., 2014). An initial proof of concept for this methodology using simulated data indicated that these methods were more robust to the DLM data constraints and provided a methodologically-sound evaluation of model fit.

The present paper builds on this research, providing model fit results for the DLM assessment using the Bayesian approach to model estimation and posterior predictive model checks. The following sections provide a brief description of the models that were estimated and the methodology used to evaluate model fit using both absolute and relative indices. Results are summarized for the 1,377 linkage levels measured by the assessment, which includes 740 linkage levels for English language arts (ELA) based on 148 EEs, 535 linkage levels for mathematics based on 107 EEs, and 102 linkage levels for science based on 34 EEs[1].

2. Description of Models

DLM assessments are currently calibrated and scored at the linkage level using a fungible log-linear diagnostic classification model (LCDM; Henson, Templin, & Willse, 2009), which in the case of a single attribute is equivalent to a fungible latent class model (Bartholomew et al., 2011). Thus, for each linkage level, it is important to assess how well the model fits the data. For each linkage level, three types of parameters are estimated: conditional probabilities of answering the items correctly for non-masters, conditional probabilities of answering the items correctly for masters, and a structural parameter that defines the base rate of mastery (i.e., the probability of randomly drawing a master from the population). Conditional probabilities represent the probability of an individual providing a correct response to the item, given that the model has classified the individual in the given mastery class.

To evaluate model fit for DLM assessments, five models were fit to each linkage level: fungible, equivalent slopes, partial equivalency with fixed variance, partial equivalency with estimated variance, and non-fungible. The fungible model assumes each item measures the linkage level equivalently, consistent with the conceptual approach to item writing. The equivalent slopes model assumes that the items are independent of each other for non-masters, but the increase in log-odds of providing a correct response is equivalent for masters on all items measuring the linkage level. The partial equivalency model does not assume equivalent parameters for all items measuring the linkage level, but rather that all item parameters come from a distribution of possible item parameters. Thus, there is some level of fungibility, or equivalency, among the items, but the amount can be estimated and varies between linkage levels. Finally, the non-fungible model makes no assumptions about equivalent item parameters, and therefore provides the most flexible estimation.

All of the models are defined similarly to a latent class model with random effects and two possible classes: masters and non-masters. Under all models, the probability of respondent j providing a correct response to item i is defined as seen in equation (1), where α_j is a binary indicator of the mastery status for respondent j.

$$P(y_{ji} = 1|\alpha_j) = \frac{\exp(\beta_0 + b_{0i} + (\beta_1 + b_{1i})\alpha_j)}{1 + \exp(\beta_0 + b_{0i} + (\beta_1 + b_{1i})\alpha_j)} \tag{1}$$

Equation (1) shows the similarity to multilevel models. In this model, β_0 and β_1 represent the attribute-level intercept and main effect respectively. These are akin to the weighted average intercept and main effect for all items measuring the linkage level (i.e., the fixed effects in the multilevel model literature). In addition to the attribute-level parameters, there are also item-level intercepts (b_{0i}) and main effects (b_{1i}). These parameters represent each item's deviation from the attribute-level effect. Thus, the full intercept for item one would be calculated as $\beta_0 + b_{01}$. This is similar to the estimation of random intercepts and slopes for each item. The difference between the proposed models and multilevel models is the treatment of the variance of these item-level parameters. In multilevel models, the variance of these random effects would be estimated. However, the variance of the random effects can also be fixed to pre-specified values.

[1] Science has three linkage levels for each EE: Initial, Precursor, and Target

If all item-level parameters are constrained to be zero, then all items will have parameters equal to the attribute-level parameter (i.e., all of the b_{0i} and b_{1i} parameters in equation (1) would be zero). This is mathematically equivalent to the fungible model. Alternatively, the item-level parameters can be allowed to vary freely with no constraints. This is mathematically equivalent to the non-fungible model. The equivalent slopes and partial equivalency models fall in the middle, using a mix of constrained and free intercepts and slopes:

- ***Fungible.*** In the fungible model, all item-level effects are fixed to 0. Thus, item level effects are not estimated.
- ***Equivalent Slopes.*** In the equivalent slopes model, the item-level slopes are fixed to be 0, but the intercepts are allowed to vary freely. Thus, we estimate no item-level main effects, only the attribute-level main effect. In contrast, we estimate no attribute-level intercept, but rather only the item-level intercepts that are allowed to vary from one another.
- ***Partial Equivalency–Fixed Variance.*** In the partial equivalency model, both attribute-level and item-level parameters are estimated. In this model, the variance of the item-level effects around the attribute-level effect is fixed. Specifically, the item-level parameters are defined to come from a $\mathcal{N}(\mu = 0, \sigma = 1)$ distribution.
- ***Partial Equivalency–Estimated Variance.*** This is similar to the partial equivalency model with fixed variance, except that the variance of the item-level effects is estimated from the data. This is a hierarchical model where the variance of the prior for the item-level effects is another estimated parameter. Thus, the prior for both the item-level intercepts and main effects are defined as $\mathcal{N}(\mu = 0, \sigma = \nu)$, where ν is an estimated parameter.
- ***Non-fungible.*** In the non-fungible model, all item-level effects are allowed to vary freely. Thus, attribute-level effects are removed from the model and only the freely estimated item-level effects are included.

In addition to inclusion/exclusion of the attribute- and item-level parameters, prior distributions must also be defined for all included parameters and the structural parameter that represents the overall base rate of mastery. For intercept parameters, a $\mathcal{N}(\mu = 0, \sigma = 2)$ prior is used. The prior was chosen as >99 percent of this distribution encompasses the plausible values for these parameters. Specifically, 99 percent of this distribution covers the log-odds range of -5.15 to 5.15, which covers nearly all of the probability scale when other parameters are equal to zero, as seen in Figure 1. The main effect parameters use a lognormal prior, Lognormal$(\mu = 0, \sigma = 1)$. These parameters are constrained to be positive to ensure monotonicity of the model. Similar to the prior for intercepts, this distribution was chosen as >99 percent of the distribution covers the range of plausible values. Specifically, this covers the log-odds range of 0 to 10.24. An upper end of ~10 was desired as a main effect of 10 would allow for an estimated probability of success near 1.0 in the extreme case where the intercept was -5 (the lower end of the intercept prior distribution). The lognormal distribution also ensures that all main effect parameters are positive. In the partial equivalency models where both attribute- and item-level effects are estimated for the same parameters (e.g., attribute- and item-level slope), the attribute-level effects use the prior distributions specified here, and the item-level effects use the fixed or estimated variance priors defined above. Finally, the structural parameter, η is defined with a flat beta distribution B$(\alpha = 1, \beta = 1)$.

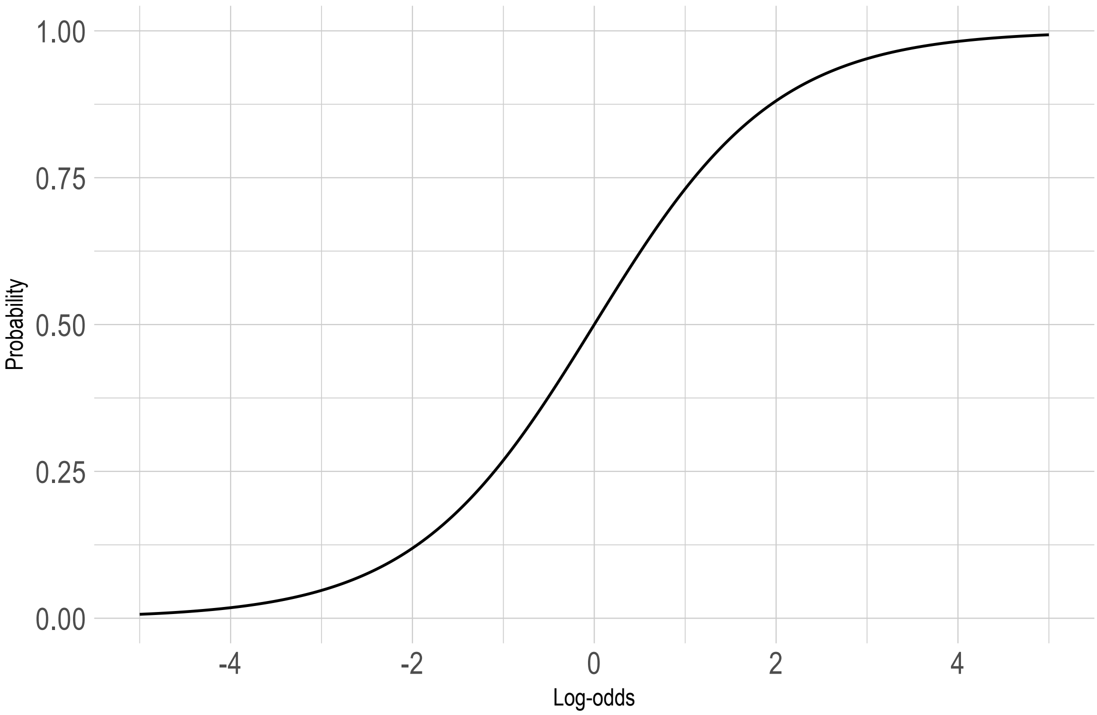

Figure 1. Log-odds to probability conversion.

3. Model Fit Calculation Background

To provide evidence of model fit for competing models (e.g., fungible, equivalent slopes, partial equivalency with fixed variance, partial equivalency with estimated variance, and non-fungible), model fit evidence can be provided in the form of both absolute and relative fit indices. Absolute fit indices evaluate how well the model fits the data. Relative fit indices compare models to each other, and are only able determine if one model provides better fit to the data, relative to the fit provided by the other model. Relative fit indices also make an assumption that the models in the comparison have acceptable absolute fit.

3.1. Absolute Fit

Absolute fit is assessed through posterior predictive model checks. Posterior predictive checks involve simulating replications of the data using the values of the posterior distributions, and then comparing the replicated data sets back to the observed data (Gelman et al., 2014). Because the replicated data sets are simulated from the current values of the parameters at each iteration of the Markov Chain, these replicated data sets represent what the data would be expected to look like *if the specified model were true*. Therefore, summaries of these data sets can be used to look for systematic differences in the characteristics of the observed data and the replicated data sets, often through visualization (Gelman & Hill, 2006).

For DLM data, there are two major posterior predictive checks that are implemented: item level p-values and raw score distributions. For the item-level analysis, the p-value is calculated for each item in each of the replicated data sets. This distribution is then compared to the p-values that were actually observed (see Figure 2). Using this check, model fit at the item level can be assessed by looking at the percent of items that have an observed p-value outside of the expected interval.

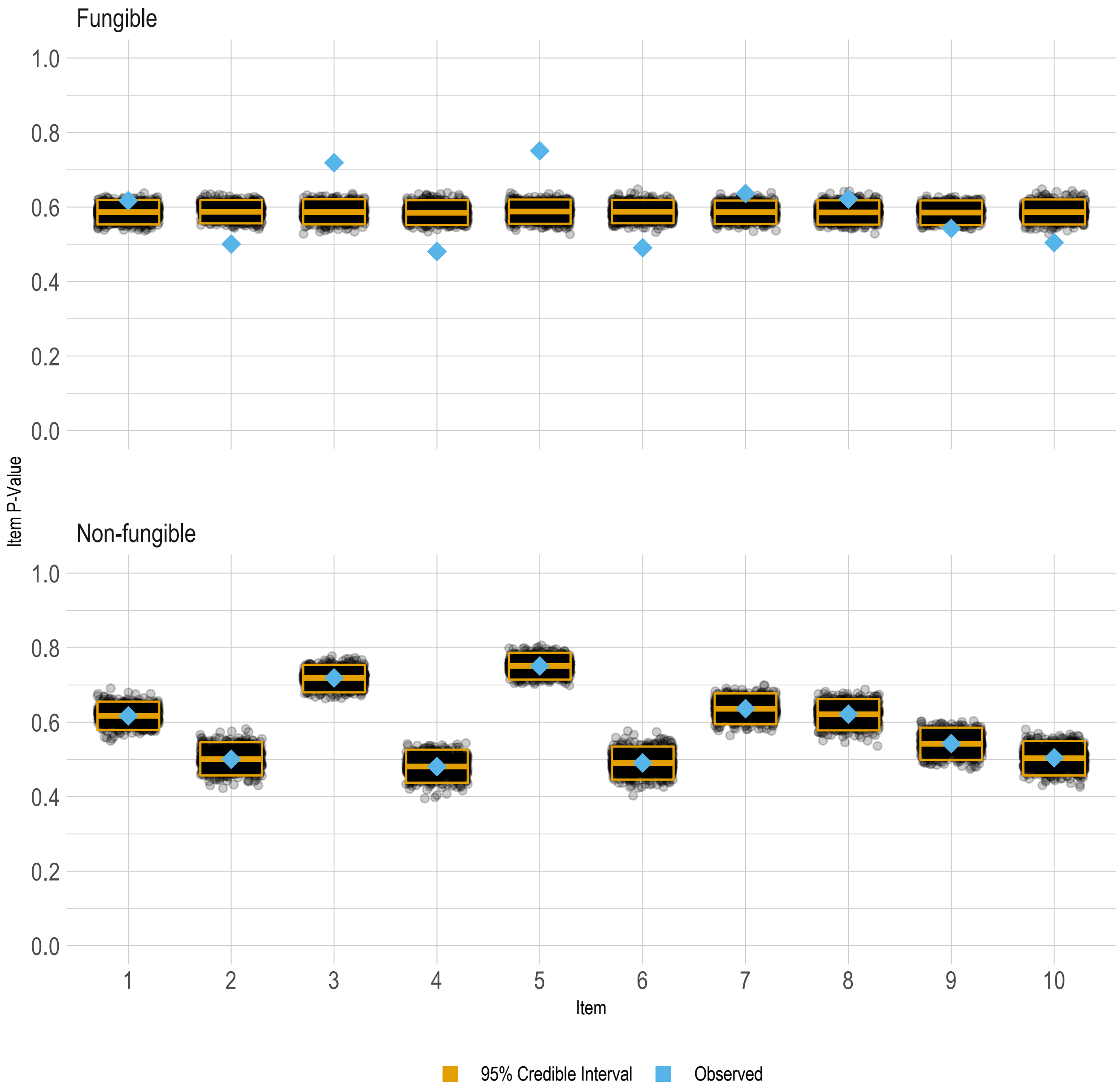

Figure 2. Example posterior checks for item-level p-values using simulated data.

At the attribute-level, model fit is assessed using the raw score distribution. That is, in each replicated data set the number of students at each raw score point can be counted. This is then compared to the number that were actually observed at the score point (see Figure 3). Based on the expected number of students at each score point, a χ^2 statistic can be calculated to determine if there is a significant amount of misfit. This χ^2 is calculated for each replicated data set and the observed data set, meaning that a posterior predictive p-value is calculated, rather than a traditional p-value. Thus, this test does not rely on the asymptotic assumptions of the traditional χ^2, which was a major limitation of the model fit analysis that used limited information tests of model fit, rather than posterior predictive model checks. Finally, unlike the item-level posterior checks, the attribute-level check captures a more complex evaluation of the data. Thus, this check offers a more robust measure of overall model fit.

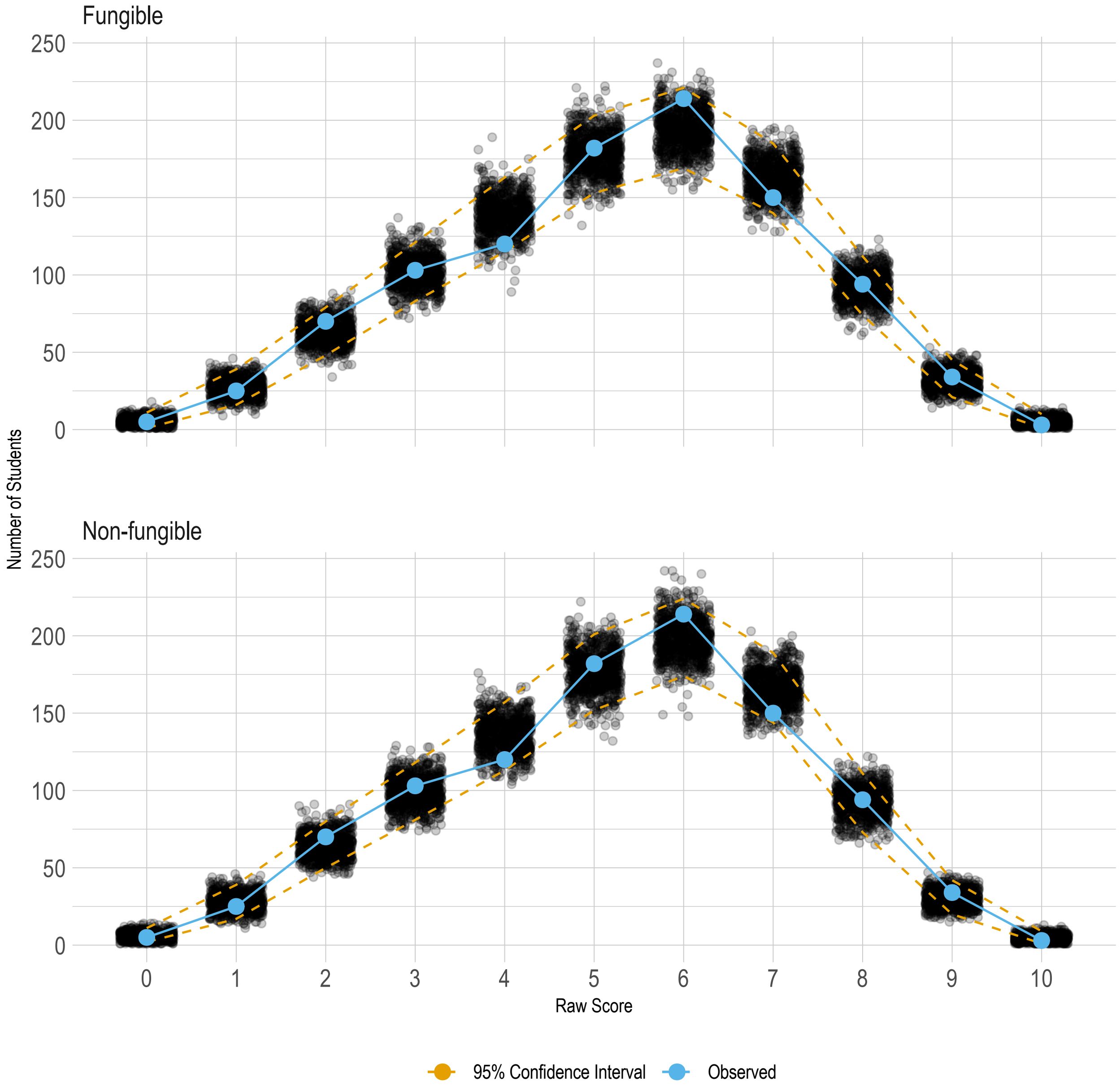

Figure 3. Example posterior predictive model check for attribute-level raw score distributions using simualted data.

3.2. Relative Fit

Relative fit is assessed through the comparison of models to determine if one model has relatively better fit than another. For DLM, two methods of comparison are used: Pareto smoothed importance sampling leave-one-out cross validation (PSIS-LOO; Vehtari, Gelman, & Gabry, 2017) and the widely applicable information criterion (WAIC; Watanabe, 2010). Both measures provide point estimates for the out of sample prediction accuracy using the log-likelihood posterior distribution. However, although the PSIS-LOO and WAIC are asymptotically equivalent, Vehtari et al. (2017) found that the PSIS-LOO is more robust than the WAIC when weak priors are used (as is true for the models utilized by DLM) and when there are influential observations (e.g., a student providing an incorrect response despite a high probability of success).

4. Procedure for Evaluating Model Fit

4.1. Data

The estimation of the models used data from the 2015–2016 assessment windows and the 2016–2017 instructionally embedded window. Field test testlets and retired testlets from previous years are not included.

4.2. Method

All models were estimated using *Stan* (Carpenter et al., 2017) using **rstan** (Stan Development Team, 2018) interface in R (R Core Team, 2018). Following the estimation, absolute fit was assessed using item- and attribute-level posterior predictive model checks. Models that demonstrated acceptable levels of absolute fit (less than 80% of linkage levels rejected for poor fit) were then compared using the PSIS-LOO and WAIC.

5. Results

5.1. Convergence

Table 1 shows the convergence rates of the different models. Convergence was determined by the Rhat statistic, which measures the within chain variance relative to the between chain variance. A model was regarded to have converged if all Rhat statistics were less than 1.1 (Gelman et al., 2014). All models converged at a high rate, with the exception of the partial equivalency with estimated variances model. Because of this finding, the partial equivalency with estimated variances model is excluded from further analyses in this paper. Interestingly, ELA linkage levels had a slightly lower convergence rate than mathematics or science. This is due in large part to writing linkage levels. Of the 50 ELA models that failed to converge across the fungible, equivalent slopes, partial equivalency with fixed variance, and non-fungible models, 43 of the linkage levels were from writing EEs. This is most likely due to how item identifiers are assigned to writing items. For scoring purposes, writing items are scored at the option level. That is, each option is treated as its own item. For more information on this approach, see Chapter 3 of the *2016–2017 Technical Manual Update—Integrated Model* (Dynamic Learning Maps Consortium [DLM Consortium], 2017). However, unlike item identifiers, option identifiers change from year to year. Because the identifiers for options changes from year to year, they appear as separate items in the calibration. The result is the appearance of more items each with less data (i.e., only one year instead of three). Thus, the data matrix is far sparser for models with writing data than for other models. For the purposes of this paper, all further analyses are based only on the models that successfully converged (e.g., summaries of the partial equivalency with fixed variances model for ELA Successor linkage levels are based on the 97.3% of linkage levels that successfully converged).

Table 1. Convergence Rates for All Models

Subject and Linkage Level	Fungible (%)	Equivalent Slopes (%)	Partial – Fixed Var. (%)	Partial – Est. Var. (%)	Non-fungible (%)
English Language Arts					
Initial Precursor	99.3	98.0	97.3	82.2	96.6
Distal Precursor	98.0	99.3	95.3	60.0	96.6
Proximal Precursor	100.0	100.0	95.9	65.1	98.0
Target	98.0	100.0	98.6	43.5	98.6
Successor	100.0	100.0	97.3	16.4	99.3
Mathematics					
Initial Precursor	100.0	100.0	100.0	100.0	100.0
Distal Precursor	100.0	100.0	98.1	71.0	99.1
Proximal Precursor	99.1	100.0	100.0	72.6	100.0
Target	100.0	100.0	99.1	67.3	99.1
Successor	100.0	100.0	100.0	34.6	99.1
Science					
Initial	100.0	100.0	100.0	76.5	100.0
Precursor	100.0	100.0	97.1	41.2	100.0
Target	100.0	100.0	100.0	2.9	100.0

5.2. Absolute Fit

For each linkage level, the percent of items that were flagged for misfit and the overall raw score distribution fit for each model were calculated using the replicated posterior data sets. Table 2 shows the average percent of items within a linkage level that were flagged for misfit for each model. For example, for the ELA Initial Precursor level estimated with the fungible model, an average of 71% of items were flagged for misfit. Table 3 shows the percentage of linkage levels that were flagged for overall misfit using the raw score distributions. For example, approximately 55% of the Target linkage levels for ELA were flagged for poor model fit to the observed raw score distribution when using the fungible model.

Overall, these results show that none of the four models that were investigated and successfully converged have acceptable levels of absolute fit. Table 2 shows consistently high percentages of items flagged for model misfit across all subjects, linkage levels, and models. Given the poor results in Table 2, it is unsurprising that Table 3 shows large proportions of linkage levels being flagged for model misfit. Across subjects, linkage levels, and models, there is consistently higher than 50 percent of linkage levels flagged for misfit. However, there is a clear trend of model fit improving as fungibility is removed from the model, as would be expected. Further, the results here compare favorably to the model fit results using the E-M algorithm that were in the initial model fit investigation that utilized limited information model fit indices. Additionally, there is a clear pattern of model fit increasing at the higher linkage levels (i.e., moving from Initial Precursor to Successor).

Table 2. Average Percent of Items Flagged for Misfit

Subject and Linkage Level	Fungible (%)	Equivalent Slopes (%)	Partial – Fixed Var. (%)	Non-fungible (%)
English Language Arts				
Initial Precursor	70.5	67.2	61.5	49.9
Distal Precursor	75.3	77.1	68.1	51.3
Proximal Precursor	76.1	88.1	76.9	54.1
Target	70.8	89.8	80.6	52.0
Successor	55.7	61.3	55.8	25.4
Mathematics				
Initial Precursor	82.0	77.5	69.6	66.5
Distal Precursor	80.7	79.3	59.4	40.4
Proximal Precursor	85.3	88.9	73.4	49.8
Target	81.4	70.7	73.0	41.5
Successor	42.0	25.8	25.8	15.8
Science				
Initial	99.0	99.4	99.0	98.6
Precursor	84.9	56.4	44.8	28.7
Target	86.6	99.1	96.5	94.1

Table 3. Percent of Linkage Levels Flagged for Model Misfit

Subject and Linkage Level	Fungible (%)	Equivalent Slopes (%)	Partial – Fixed Var. (%)	Non-fungible (%)
English Language Arts				
Initial Precursor	70.5	67.2	61.5	49.9
Distal Precursor	75.3	77.1	68.1	51.3
Proximal Precursor	76.1	88.1	76.9	54.1
Target	70.8	89.8	80.6	52.0
Successor	55.7	61.3	55.8	25.4
Mathematics				
Initial Precursor	82.0	77.5	69.6	66.5
Distal Precursor	80.7	79.3	59.4	40.4
Proximal Precursor	85.3	88.9	73.4	49.8
Target	81.4	70.7	73.0	41.5
Successor	42.0	25.8	25.8	15.8
Science				
Initial	99.0	99.4	99.0	98.6
Precursor	84.9	56.4	44.8	28.7
Target	86.6	99.1	96.5	94.1

5.3. Relative Fit

Relative fit compares two competing models to determine which model provides better fit, relative to the other. However, one of the assumptions of these methods is that the models being compared have acceptable levels of absolute model fit. Because none of the fungible, equivalent slopes, or partial equivalency models showed sufficient absolute model data fit, the relative fit analyses were not estimated.

6. Summary of Model Fit Analyses

In this paper, five models were estimated for the DLM alternate assessment with varying levels of fungibility. Of these models, all except the partial equivalency with estimated variance showed high convergence rates. However, the remaining four models showed poor fit to the data using posterior predictive model checks. There are several reasons this may be the case. First, it is possible that fungibility is a poor assumption for this data, and any attempt to add fungibility results in poor fit to the underlying data. Although this could certainly be the case for the fungible, equivalent slopes, and partial equivalency with fixed variance models, this explanation is insufficient for the non-fungible model.

Alternatively, because students take a small number of items per linkage level, it is likely that there is a large degree of uncertainty in the resulting posterior probabilities of mastery. This would result in large differences between the replicated data sets used for the posterior predictive checks. Without a clear picture of what true model-fitting data looks like, it may be difficult for the observed data to match the distribution of posterior data sets, leading to poor model fit.

It is also important to note that these analyses cannot answer the question of practical significance. Given the high flagging rate for the non-fungible model, it is possible that these methods are too sensitive to small violations of model fit. Thus, although there is statistically significant model misfit, the practical implications may be negligible. This hypothesis is supported by preliminary analyses using the 2016–2017 operational assessment, which showed minimal differences between student assessment scores derived from the fungible and nonfungible models. Further work is needed to investigate the potential impacts of different scoring models.

Several simulation studies are planned to investigate these potential causes in order to further refine the methodology for evaluating the model fit of the DLM system. Specifically, current work is focused on evaluating the properties of the various models and the model fit indices. For example, data can be simulated from the non-fungible model and then estimated with both the fungible and non-fungible model. The agreement between the mastery profiles from the two models can provide some measure of the practical significance of model misfit for this type of assessment. In another study we are examining the sensitivity and specificity of various model fit measures for diagnostic assessments to more fully understand which methods are likely to provide the most valid inferences about model fit. Additionally, future work is planned to identify items that are exhibiting misfit to look for patterns in the items that may explain why the items exhibit misfit, and may also inform future test development.

Finally, although the analyses presented here are specific to the DLM alternate assessment, these findings and lessons learned can be applied to almost any diagnostic and/or adaptive assessment. For example, finding an adequate measure of model-level fit with a sparse data matrix is applicable to any adaptive test where students don't test on all of same items, or even necessarily sets of items. This work also has implications for diagnostic assessments, where the evaluation of model is traditionally limited to only model comparisons, without also evaluating the absolute fit of the model to the data.

References

Bartholomew, D., Knott, M., & Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*. West Sussex, UK: Wiley.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). doi:10.18637/jss.v076.i01[2]

Dynamic Learning Maps Consortium. (2017). *2016–2017 Technical Manual Update—Integrated Model*. University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS). Lawrence, KS.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis* (3rd). Boca Raton, FL: CRC Press.

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models* (1st). Cambridge, England: Cambridge University Press.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210. doi:10.1007/s11336-008-9089-5[3]

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732. doi:10.1007/s11336-005-1295-9[4]

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*(4), 305–328. doi:10.1080/00273171.2014.911075[5]

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Stan Development Team. (2018). RStan: The R interface to Stan. R package version 2.17.3. Retrieved from http://mc-stan.org/

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[6]

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.

[2] https://dx.doi.org/10.18637/jss.v076.i01

[3] https://dx.doi.org/10.1007/s11336-008-9089-5

[4] https://dx.doi.org/10.1007/s11336-005-1295-9

[5] https://dx.doi.org/10.1080/00273171.2014.911075

[6] https://dx.doi.org/10.1007/s11222-016-9696-4