DYNAMIC®
LEARNING MAPS

# *Using Propensity Scores to Evaluate Changes in Cross-Year Performance Distributions*

Technical Report #21-01

**December 2021**

# Contents

# List of Tables

# List of Figures

# Executive Summary

For assessments that are administered annually, it is common to compare performance across administration years. However, when there are changes to the student population over time, it can be difficult to make comparisons. Regardless of whether population differences are due to an acute event (e.g., the COVID-19 pandemic), long-term systematic change (e.g., compliance with the Every Student Succeeds Act 1% threshold for alternate assessments based on alternate achievement standards [AA-AAS] participation), or random fluctuations, changes in the population need to be accounted for in order to evaluate changes in performance from year to year. Further, after accounting for changes in the population, any observed differences in performance must be evaluated to determine whether these differences are aberrant or consistent with normal year-to-year fluctuations.

This report outlines an approach for evaluating differences in performance distributions across years. Specifically, this report describes the following methods:

- the use of a propensity score model to estimate the probability that each student would participate in an administration year based on demographic characteristics and survey responses
- a matching algorithm based on random resampling that accounts for population differences while preserving the original sample size of both administration years
- a process for flagging aberrant changes based on the effect size of the change
- a procedure for visually reporting results that simultaneously communicates both the magnitude and importance (effect size) of each change

Findings illustrate the utility of the proposed methods for making cross-year comparisons. Specifically, the findings demonstrate the accuracy of the propensity score model and the effectiveness of the matching algorithm for minimizing differences in the population across administration years. Additionally, the effect size visualization method allows readers to readily identify which changes are within expected limits and which changes may need further investigation.

# 1. Background and Purpose of the Report

Dynamic Learning Maps® (DLM®) assessments report student achievement as an overall performance level for each subject, which is used for state accountability systems. There are four performance levels used to describe student achievement: *Emerging*, *Approaching the Target*, *At Target*, and *Advanced*.

When administering DLM assessments in English language arts (ELA) and mathematics, states choose between the Instructionally Embedded and Year-End models. In the Instructionally Embedded model, testlet are administered in both a fall and spring window, and teachers choose which subset of Essential Elements to assess and which linkage level is selected for each Essential Element. In the Year-End model, only assessments from the spring assessment window are included for summative scoring, and both the Essential Elements and linkage levels system-assigned. All states participating in science follow a single administration model that resembles the Year-End ELA and mathematics model, regardless of which model is selected for ELA and mathematics.

After the close of each administration year, DLM staff summarize student performance level distributions within each assessment model (Instructionally Embedded and Year-End), grade, and subject (ELA, mathematics, and science). In isolation, these numbers can be difficult to interpret. To add context to the raw percentages of students who achieved at each performance level, DLM staff also report the change in each performance level from the prior year. This additional context can help state partners in interpreting the results from a given year.

However, the change in performance level distributions can be misleading when there has been change to the student population. Accounting for changes to the student population when making performance comparisons across years has become increasingly important in the aftermath of the COVID-19 pandemic. For example, the COVID-19 pandemic impacted whether and how students receive instruction as well as whether students are able to participate in assessments (Cui, 2020). This issue of population changes across years is not unique to evaluating COVID-19 related disruptions. There has been a systemic change to the population of students taking alternate assessments based on alternate achievement standards (AA-AAS) as states comply with the 1% participation threshold outlined in the Every Student Succeeds Act (ESSA, 2015). In the general education setting, the student population has changed over the last several years as a result of the assessment "opt-out movement," where sometimes large groups of students decide to opt-out and not complete any assessments (Bennett, 2016; Marland et al., 2019). When the population of students completing an assessment changes, it is important that the population is balanced across administration years to ensure that the comparison of performance level distributions is not simply reflecting those changes to the population.

In this report, we describe a method to account for population changes that can be applied when there have been acute disruptions, as well as long-term systematic efforts to change the student population. The method is based propensity score matching and can be used to estimate changes in performance level distributions across years, identify aberrant changes, and present the findings visually. Each of these features of the method are described in turn. Throughout the report we use the 2017–2018 and 2018–2019 administration of the DLM assessments as an illustrative example. That is, we estimate the change in the performance level distributions between the 2017–2018 and 2018–2019 administration years. Although we focus on the application of the proposed methods to DLM assessments, psychometricians in other assessment programs can use this report as a guide for building their own

models to evaluate year-to-year differences in performance levels.

# 2. Estimate Change in Performance Level Distributions

To make accurate comparisons of the performance level distributions for a grade and subject across years, we consider whether the samples from the two years are consistent. In other words, we want to be sure that any changes we observe in the distributions from Year $X-1$ to Year $X$ are due to actual changes in performance rather than changes to the student population. As an example, if the highest performing students exited from the DLM assessment as states complied with the ESSA 1% threshold, the overall distribution may look like performance declined. However, this would most likely be due to a change in the population, rather than an actual decline in performance. Therefore, it is important to balance the samples across years to make accurate comparisons.

To balance the samples, we employ a propensity score model. Propensity scores represent the probability of an individual being assigned to a group, conditional on baseline characteristics (Austin, 2011). For this analysis, the group assignment is the assessment administration year. That is, given *a priori* student characteristics, can we predict which year the student was assessed? If there are significant changes to the student population across years, then we should be able to make more accurate predictions of which year a student was assessed in than if the populations are consistent across years. Conversely, if the student population is perfectly stable across years, our predictions should not be any more accurate than chance.

After considering several methods for estimating the propensity scores, we recommend implementing a random forest model (Wright & Ziegler, 2017), as this method shows the best performance for predicting administration year from student baseline characteristics (see Appendix A for details on the model selection process).

Random forests are a type of decision trees, which are logic-based algorithms where one or more predictor variables are added to each layer of the tree to split the data. Specifically, a random forest model is comprised of a multitude of trees that each contribute to the model-assigned output for each data point (Breiman, 2001; Kotsiantis, 2013). The trees in random forest models can be classification trees (i.e., decision trees), regression trees (i.e., classification and regression trees), or survival trees (Wright & Ziegler, 2017), with classification and regression trees (CARTs) being the most applicable for calculating propensity scores. The key difference between CARTs and decision trees is that CARTs produce continuous scale output, while decision trees produce nominal scale classification output (XGBoost Developers, 2020). Growth and pruning phases are used in constructing trees. The growth phase entails splitting nodes until all data points are classified perfectly or until the number of data points at each node is less than the minimum number required for adding another layer to the tree (i.e., a stopping criterion). The pruning phase entails simplifying the lower layers of the tree to reduce the risk of the tree overfitting the data, particularly if a stopping criterion was not utilized during the growth phase. The performance of individual trees is dependent on the order that the predictor variables were added to the tree; hence, random forests incorporate many trees to reduce the impact of the order that the predictor variables were added to each tree, and this typically leads to a drastic increase in the performance of a random forest compared to an individual tree (Breiman, 2001; Kubat, 2017).

## 2.1. Propensity Score Model Estimation

For each assessment model (i.e., Instructionally Embedded or Year-End), grade, and subject, a random forest is estimated, predicting administration year from a set of demographic variables from enrollment files and responses to the First Contact survey shown in Table 1. The selected demographic variables are those that are reported annually in the DLM Technical Manual (i.e., gender, race, Hispanic ethnicity, English learning [EL] status), and those from the First Contact survey that could show important differences across years (i.e., primary disability, computer use, educational placement, primary language). Additionally, responses to the First Contact survey are used to calculate subject-specific and communication complexity bands for each student. Complexity bands are used to recommend linkage levels for Instructionally Embedded ELA and mathematics assessments and assign the first linkage level for science and Year-End ELA and mathematics assessments. For more information on the contents and use of the First Contact survey, see Nash et al. (2016).

**Table 1**

*Variables Included in the Propensity Score Models*

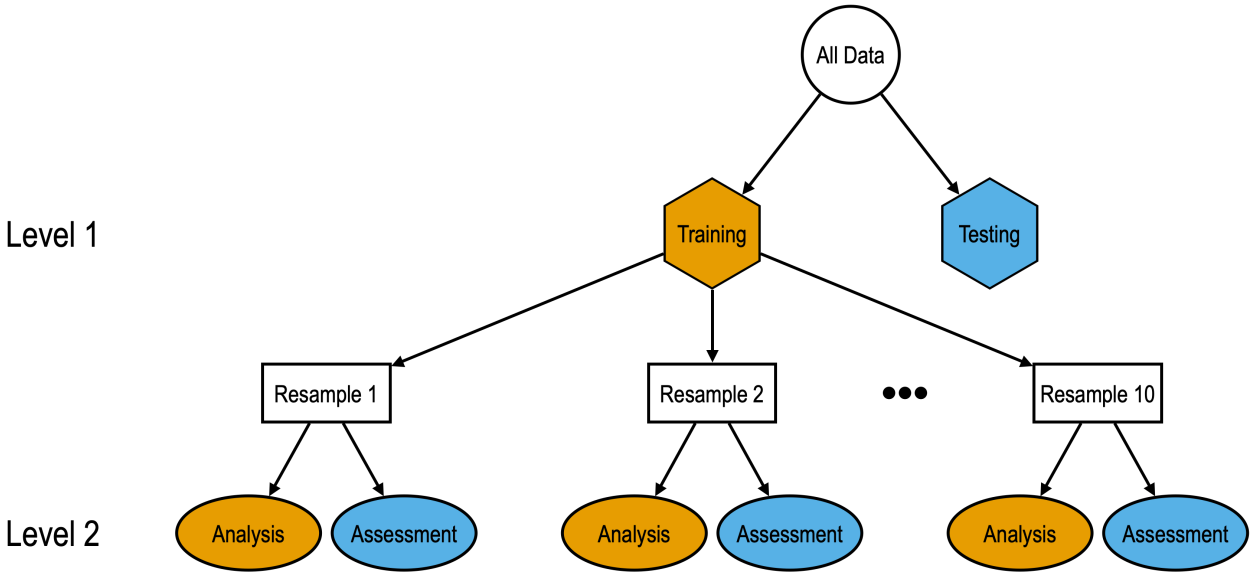
| Predictor | Description |
|---|---|
| Gender | Student's reported gender |
| Race | Student's reported race |
| Hispanic ethnicity | Student's recognition of their Hispanic ethnicity |
| Primary disability | Student's reported disability category |
| English learning (EL) participation | Student's eligibility or participation in EL services |
| Computer use | Student's primary use of a computer during instruction |
| Educational placement | Student's educational placement (e.g., regular class, resource room, separate class, etc.) |
| Primary language | Is English the student's primary language (Yes or No) |
| Subject-specific complexity band | Student's subject complexity band, based on domain-specific questions on the First Contact survey (i.e., ELA, mathematics, and science) |
| Writing complexity band* | Student's writing complexity band, based on responses to the First Contact survey |
| Expressive communication band | Student's expressive communication band, based on responses to the First Contact survey |
| Receptive communication score | Student's sum score of receptive communication items on the First Contact survey |

* Only included for English language arts models.

The random forest model also includes hyperparameters than cannot be estimated directly from the data. These include the number of predictors that are randomly sampled at each split when creating individual trees and the minimum number of data points in a node that are required for the node to split further. To select hyperparameters for the final models, a nested resampling approach is used, as illustrated in Figure 1.

**Figure 1**

*The Nested Resampling Approach*




The full data for a model, grade, and subject is initially split into a training and a testing set. This is the Level 1 split. For the initial split, 75% of the data is randomly selected for the training set, and the remaining 25% is reserved for the testing set. The training set is then resampled using v-fold (also called k-fold) cross validation (de Rooij & Weeda, 2020; Simon, 2007). Specifically, 10 folds are created. This means that the training set is partitioned into 10 analysis and assessment sets.

For each split, model performance is evaluated using a receiver operating characteristic (ROC) curve (Flach et al., 2011; Tharwat, 2020). ROC curves plot the true positive rate of a classifier (i.e., sensitivity) against the false positive rate (i.e., 1 - specificity). ROC curves that fall directly on the diagonal represent prediction equal to random guessing. ROC curves that are above the diagonal represent increased predictive accuracy, with ROC curves that reach the top left corner representing perfect predictions. We can quantify the performance of each model by calculating the total area under the ROC curve. An area of 1.0 indicates perfect prediction (i.e., the curve goes all the way up to the top left corner), and a value of .5 indicates chance guessing (i.e., the curve lies directly on the diagonal).

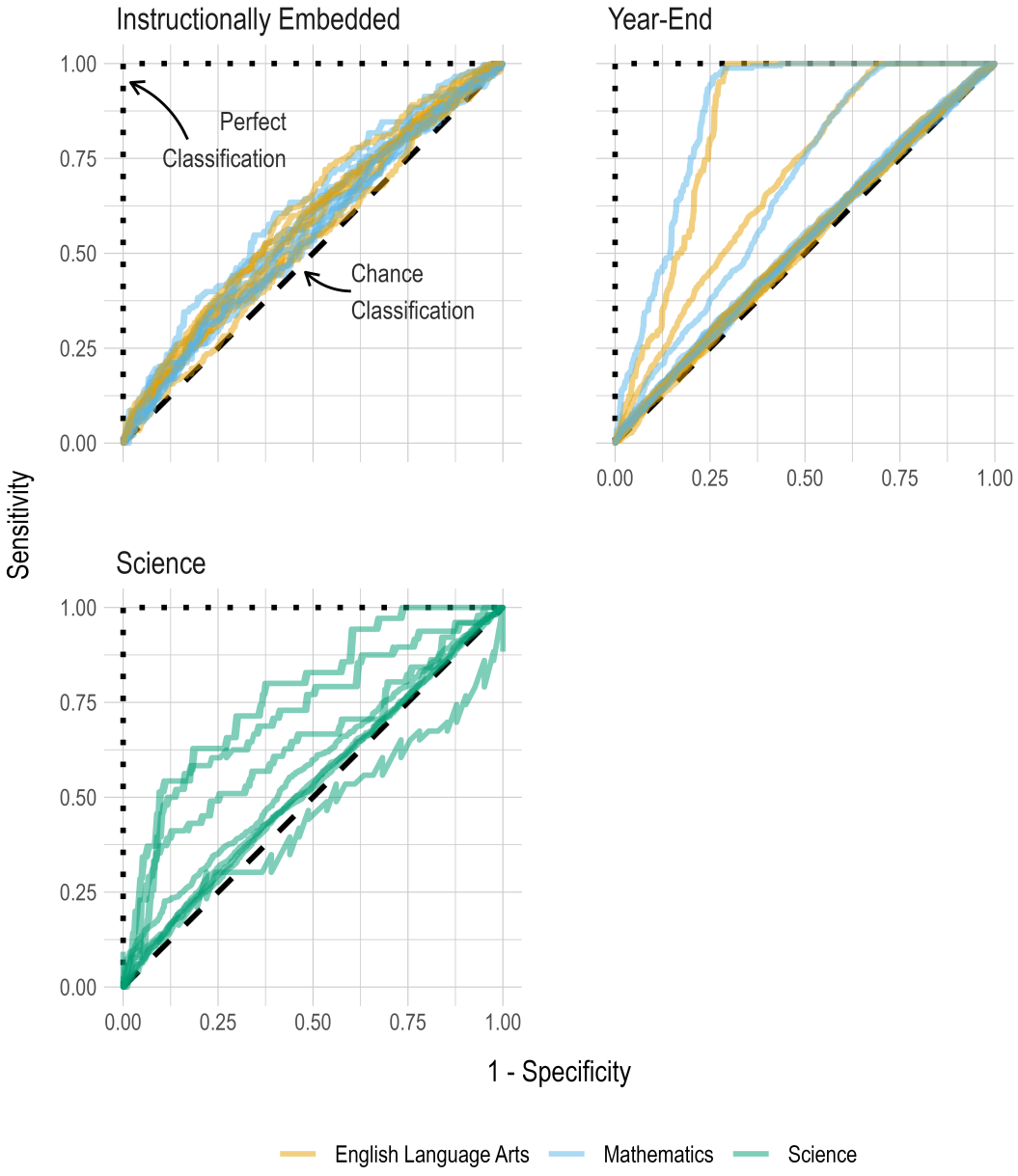
Given these data splits and this performance metric, model estimation for each assessment model, grade, and subject proceeded using the following process:

1. A model was estimated for each set of candidate hyperparameters for the random forest on each of the 10 analysis sets.
2. Model performance was evaluated by using the fitted model to predict the assessment sets and calculating the area under the ROC curve.
3. Optimal hyperparameters for random forest models were selected based on which hyperparameter(s) maximized the area under the ROC curve.
4. The selected hyperparameters were used to fit a final random forest to the complete set of training data. These models were then used to predict the test set.

Figure 2 shows the ROC curves for each model, grade, and subject when comparing the 2017–2018 administration to 2018–2019. In general, the random forest model predicts slightly better than chance, with ROC curves slightly above the diagonal. This is also reflected in Table 2, which reports the area under each of the ROC curves. The area under each ROC curve ranges from .45 to .87, with a median of .54. This indicates that the majority models showed better than chance accuracy in predicting administration year, but also that none of the models had notably high prediction accuracy.

**Figure 2**

*ROC Curves for the Implemented Propensity Score Models*



Ideally, the student population would be perfectly stable across years, and our propensity score model would not be able predict administration year any better than chance. We know from feedback from state partners that this is not the case. For example, many states participating in the DLM assessments reported that the student population did have changes between 2017–2018 and 2018–2019 as a result of ESSA compliance. Thus, it is expected and a positive finding that the propensity score model is able to predict better than chance. This indicates that the model is able to detect small differences in the population. In summary, the ROC curves indicate that the samples are largely consistent, as the predictive accuracy is not much larger than chance, but that there are small differences in the samples that should be accounted for when making comparisons.

**Table 2**

*Area Under the ROC Curve, by Assessment Model, Grade, and Subject*

| Grade or course | English language arts | Mathematics | Science |
|---|---|---|---|
| Instructionally Embedded | | | |
| 3 | .535 | .542 | —[†] |
| 4 | .578 | .562 | —[†] |
| 5 | .539 | .575 | —[†] |
| 6 | .567 | .590 | —[†] |
| 7 | .568 | .551 | —[†] |
| 8 | .589 | .554 | —[†] |
| 9 | .617 | .611 | —[†] |
| 10 | —[*] | .598 | —[†] |
| 11 | .566 | .541 | —[†] |
| Year-End | | | |
| 3 | .529 | .520 | .778 |
| 4 | .515 | .517 | .570 |
| 5 | .526 | .529 | .534 |
| 6 | .516 | .539 | .719 |
| 7 | .526 | .526 | .634 |
| 8 | .534 | .525 | .524 |
| 9 | .693 | .668 | .530 |
| 10 | .840 | .867 | —[‡] |
| 11 | .516 | .536 | —[‡] |
| Biology | — | — | .449 |

[*] Instructionally embedded ELA assessments are grade banded for grades 9–10 and 11–12.

[†] All science assessments follow a year-end assessment model.

[‡] Science assessments are grade banded for grades 9–12.

## 2.2. Propensity Score Matching

Once the random forest model has been trained, we can predict which administration year a student belongs to by their baseline characteristics. This results in a probability that each student would have been assessed in each administration year. These probabilities can then be used to balance the samples across years. The goal of these analyses is to determine the amount of change in the performance levels distribution in Year X compared to Year X – 1. Therefore, we must first ask what the performance level distributions in Year X – 1 would be if the sample in Year X – 1 looked like the sample in Year X. That is, we want to examine what the performance level distributions in Year X – 1 would have looked like if all factors related to the educational and assessment experience of students were the same, but the population of students more closely resembled the student population for Year X. To do this, we resample with replacement for the Year X – 1 data, weighted by the probability that the student belongs to the Year X sample. Thus, students that more closely resemble the Year X sample are more likely to be selected, meaning that the resampled data should approximate the characteristics of the Year X data.

This approach to balancing the samples was chosen for two main reasons. First, for reporting purposes, it is desirable for raw Year X performance level distributions to match the distributions used for the year-to-year comparisons, since the raw distributions and year-to-year changes are reported together. This is why we only resample the Year X – 1 data. The Year X data remains as-is to maintain consistency across the two analyses. Second, this method preserves the original sample size from each administration year. This is important because the sample size influences the amount of uncertainty around the percentage of students at each performance level. Traditional propensity score matching algorithms attempt to create a 1-to-1 match between the two samples (Caliendo & Kopeinig, 2008; Powell et al., 2020). However, for this particular analysis, increasing or decreasing the sample size may artificially decrease or increase the uncertainty of our estimated performance level distributions. Rather than matching individual students, we are interested in matching overall distributions of demographic variables. Thus, we adopt our own pseudo-matching algorithm.

Table 3, Table 4, and Table 5 show the makeup of the matched samples by demographic, First Contact survey, and complexity band variables, respectively, when the matching algorithm is applied to the 2017–2018 and 2018–2019 data. For each variable, the percentage of students in each category is reported for the raw 2017–2018 data and the matched 2017–2018 data created with the resampling process. The tables also include differences between the percentage that was observed in 2018–2019 and each of the raw and adjusted percentages from 2017–2018. If the propensity score model and the matching algorithm are functioning as expected, the difference between 2018–2019 and the adjusted sample should be smaller in magnitude than the difference between 2018–2019 and the raw sample from 2017–2018. That is, the adjusted sample should look more similar to 2018–2019 than the original 2017–2018 sample.

**Table 3**

*Distribution of Demographic Variables in the Propensity Score Models*

| Variable | 2017–2018 (%) raw | 2017–2018 (%) adjusted | 2018–2019 (%) | Raw difference | Adjusted difference |
|---|---|---|---|---|---|
| Gender | | | | | |
| Male | 67.1 | 67.1 | 66.9 | −0.2 | −0.2 |
| Female | 32.9 | 32.9 | 33.1 | 0.2 | 0.2 |
| Race | | | | | |
| White | 60.6 | 61.1 | 60.2 | −0.4 | −0.8 |
| African American | 21.9 | 21.9 | 21.6 | −0.2 | −0.2 |
| Two or more races | 8.9 | 8.7 | 10.1 | 1.1 | 1.4 |
| Asian | 5.1 | 5.0 | 4.9 | −0.2 | −0.1 |
| American Indian | 2.8 | 2.7 | 2.5 | −0.3 | −0.2 |
| Native Hawaiian or Pacific Islander | 0.5 | 0.4 | 0.5 | 0.0 | 0.0 |
| Alaska Native | 0.3 | 0.3 | 0.2 | −0.1 | −0.1 |
| Hispanic ethnicity | | | | | |
| Non-Hispanic | 78.8 | 79.4 | 79.5 | 0.7 | 0.2 |
| Hispanic | 21.2 | 20.6 | 20.5 | −0.7 | −0.2 |
| Primary disability | | | | | |
| Autism | 27.9 | 27.9 | 27.4 | −0.5 | −0.5 |
| Intellectual disability | 24.7 | 25.1 | 24.6 | −0.1 | −0.5 |
| Eligible individual | 19.0 | 19.2 | 21.4 | 2.4 | 2.2 |
| Multiple disabilities | 13.9 | 13.7 | 13.1 | −0.8 | −0.6 |
| Other health impairment | 4.7 | 4.7 | 4.8 | 0.2 | 0.1 |
| Speech or language impairment | 2.2 | 2.1 | 2.0 | −0.2 | −0.1 |
| Documented disability | 2.2 | 2.0 | 2.0 | −0.3 | −0.1 |
| Specific learning disability | 1.9 | 1.8 | 1.5 | −0.3 | −0.3 |
| Developmentally delayed | 0.9 | 0.8 | 1.0 | 0.1 | 0.2 |
| Emotional disturbance | 0.7 | 0.7 | 0.5 | −0.1 | −0.2 |
| Traumatic brain injury | 0.5 | 0.4 | 0.5 | 0.0 | 0.0 |
| Decline to answer | 0.3 | 0.3 | 0.3 | 0.1 | 0.1 |
| Orthopedic impairment | 0.4 | 0.4 | 0.3 | 0.0 | −0.1 |
| Hearing impairment | 0.3 | 0.3 | 0.3 | 0.0 | 0.0 |
| Visual impairment | 0.3 | 0.3 | 0.2 | 0.0 | 0.0 |
| No disability | 0.3 | 0.3 | 0.1 | −0.3 | −0.2 |
| Deaf/blindness | 0.1 | 0.1 | 0.1 | 0.0 | 0.0 |
| English learner (EL) participation | | | | | |
| Not EL eligible or monitored | 93.3 | 93.7 | 93.7 | 0.4 | 0.0 |
| EL eligible or monitored | 6.7 | 6.3 | 6.3 | −0.4 | 0.0 |

**Table 4**

*Distribution of First Contact Survey Variables in the Propensity Score Models*

| Variable | 2017–2018 (%) raw | 2017–2018 (%) adjusted | 2018–2019 (%) | Raw difference | Adjusted difference |
|---|---|---|---|---|---|
| Computer use | | | | | |
| Cannot access a computer | 4.2 | 4.1 | 4.2 | 0.0 | 0.1 |
| No opportunity to access a computer | 1.8 | 1.7 | 1.9 | 0.1 | 0.2 |
| Uses with human support | 51.2 | 51.1 | 51.3 | 0.1 | 0.3 |
| Accesses with assistive technology | 3.1 | 3.1 | 3.1 | 0.0 | 0.0 |
| Accesses independently | 39.7 | 40.0 | 39.5 | −0.2 | −0.5 |
| Instructional setting | | | | | |
| Homebound/hospital | 0.5 | 0.5 | 0.5 | 0.0 | 0.0 |
| Residential facility | 0.9 | 0.9 | 0.8 | −0.1 | −0.1 |
| Separate school | 29.0 | 28.3 | 27.4 | −1.5 | −0.9 |
| Less than 40% of day in regular class | 52.8 | 53.8 | 55.4 | 2.6 | 1.6 |
| 40-79% of day in regular class | 12.7 | 12.7 | 12.5 | −0.3 | −0.2 |
| 80% of day or more in regular class | 4.2 | 3.8 | 3.5 | −0.7 | −0.3 |
| English primary language | | | | | |
| Yes | 92.8 | 93.0 | 92.6 | −0.2 | −0.4 |
| No | 7.2 | 7.0 | 7.4 | 0.2 | 0.4 |

**Table 5**

*Distribution of Complexity Band Variables in the Propensity Score Models*

| Variable | 2017–2018 (%) raw | 2017–2018 (%) adjusted | 2018–2019 (%) | Raw difference | Adjusted difference |
|---|---|---|---|---|---|
| Communication band | | | | | |
| Foundational | 7.3 | 7.2 | 7.3 | 0.0 | 0.2 |
| Band 1 | 22.4 | 22.5 | 22.9 | 0.5 | 0.5 |
| Band 2 | 23.5 | 23.6 | 24.2 | 0.7 | 0.6 |
| Band 3 | 46.7 | 46.7 | 45.5 | −1.2 | −1.2 |
| English language arts band | | | | | |
| Foundational | 11.4 | 11.4 | 11.8 | 0.4 | 0.4 |
| Band 1 | 32.9 | 32.8 | 34.0 | 1.1 | 1.2 |
| Band 2 | 40.7 | 40.4 | 40.1 | −0.5 | −0.3 |
| Band 3 | 15.0 | 15.4 | 14.0 | −1.0 | −1.4 |
| Mathematics band | | | | | |
| Foundational | 12.1 | 12.0 | 12.2 | 0.2 | 0.3 |
| Band 1 | 35.6 | 35.7 | 36.5 | 0.9 | 0.8 |
| Band 2 | 40.8 | 40.8 | 40.8 | 0.0 | 0.0 |
| Band 3 | 11.6 | 11.6 | 10.5 | −1.1 | −1.1 |
| Science band | | | | | |
| Foundational | 14.3 | 14.2 | 14.2 | −0.1 | 0.0 |
| Band 1 | 41.0 | 41.0 | 41.8 | 0.8 | 0.8 |
| Band 2 | 30.0 | 30.0 | 30.4 | 0.4 | 0.4 |
| Band 3 | 14.7 | 14.8 | 13.6 | −1.1 | −1.2 |
| Receptive communication* | 18.1 | 18.1 | 18.0 | −0.1 | 0.0 |

* Values for the receptive communication scale represent the average score.

To evaluate the effectiveness of the matching algorithm, we can calculate the mean absolute difference between the 2018–2019 percentages and both the raw and adjusted 2017–2018 percentages. In addition, we can calculate a weighted mean absolute difference, where larger groups are given more weight relative to groups with smaller counts. If the matching algorithm works as expected, we would expect to see smaller differences when using the adjusted percentages (i.e., the sample is more similar to 2018–2019). Table 6 shows the unweighted and weighted mean absolute differences for each type of variable. For example, when using the raw 2017–2018 sample, there was a mean absolute difference of 0.34 percentage points from 2017–2018 to 2018–2019 across all demographic variables (Table 3). When using the adjusted sample, this differences dropped to 0.29. Similarly, the weighted mean absolute difference was 0.49 when using the raw data from 2017–2018, compared to only 0.38 when using the adjusted sample.

Overall, across all variables, there was a mean absolute difference of 0.44 and a weighted mean absolute difference of 0.60 between the 2018–2019 distributions and the raw 2017–2018 distributions. When using

the adjusted sample for 2017–2018, these values decreased to 0.40 (unweighted) and 0.52 (weighted). Thus, overall and by variable subsets, the propensity score model and matching algorithm succeed in making the 2017–2018 sample more closely resemble the 2018–2019 sample. Given these findings, we can proceed with estimating change in the performance level distributions using the adjusted sample from 2017–2018.

**Table 6**

*Mean Absolute Difference From Raw and Adjusted 2017–2018 Distributions to 2018–2019*

| Variable set | Unweighted raw | Unweighted adjusted | Weighted raw | Weighted adjusted |
|---|---|---|---|---|
| Demographic | 0.341 | 0.287 | 0.494 | 0.378 |
| First Contact | 0.463 | 0.385 | 0.755 | 0.627 |
| Complexity Band | 0.597 | 0.618 | 0.620 | 0.612 |

# 3. Identifying Aberrant Changes

Once we have our two samples (the current year's data and the resampled data from the comparison year), we can calculate the percentage of students achieving at each performance level within each grade and subject. However, these percentages are estimates of the true percentage. That is, if the entire assessment were delivered repeatedly, slightly different percentages would be observed for each repetition. The amount of uncertainty in each percentage can be quantified through the estimation of the standard error, and the standard errors from the two comparison years can be used to estimate a standard error of the change in percentage across years.

The standard error is a measure of uncertainty around the estimated percentage or change. The smaller the standard error, the more certainty there is in the estimate. The standard errors of the percentages are estimated using a multinomial log-linear model. Although standard errors could be calculated for each of the estimated percentages as if they were binomial percentages (e.g., Agresti & Coull, 1998; Brown et al., 2001), this is not strictly accurate. The multinomial model yields the same estimated percentages but is able to account for non-binary categories when estimating the standard errors. The model is estimated by predicting the counts of each performance level from an intercept and a categorical indicator of the administration year.

$$\log(n_c) = \beta_0 + \beta_1 X_{year} \quad (1)$$

The model parameters are then used to calculate the estimated percentages and standard errors using estimated marginal means (Harvey, 1960; Searle et al., 1980). Estimated marginal means are useful for evaluating the true impact of different features (i.e., administration year) on an outcome variable (i.e., performance level classifications) when there are unequal subclass sizes. This is often the case for DLM assessments, where sample sizes vary between years, and the percentages of students in each performance level are unequal. Standard errors are calculated for the estimated marginal means,

providing the best possible approximation of the uncertainty around an observed percentage of students at a given performance level.

Finally, just as with standard model estimates, pairwise contrasts can be calculated to estimate the change in percentages across years using the estimated marginal means. The contrasts estimate the change from the yearly percentages and the standard error of the change from the associated yearly standard errors. As with the model estimates, the change and standard errors account for unequal subclass sizes. The pairwise contrasts allow for multiple interpretations. First, the estimated marginal means can be used to calculate a *t*-statistic and *p*-value for the change. Because multiple comparisons are made for each model (i.e., multiple performance levels), a Tukey correction is applied. However, statistical significance does not necessarily mean that the changes are practically important. Thus, two effect sizes can be calculated. This first effect size is Cohen's *d* (Cohen, 1988). Cohen's *d* is calculated as

$$d = \frac{\mu_1 - \mu_2}{\sqrt{(\sigma_1^2 + \sigma_2^2) \div 2}} \tag{2}$$

where $\mu_1$ and $\mu_2$ are estimated mean (i.e., percentage) in each year, and $\sigma_1^2$ and $\sigma_2^2$ represent the squared standard errors for the estimates. Thus, Cohen's *d* represents a standardized mean difference. However, Cohen's *d* does make an implicit assumption that the means being compared are continuous. This is not true for performance level distributions, which are represented by percentages, and therefore are bound between 0 and 1. Thus, Cohen's *d* can be misleading, especially in cases where the estimated percentage is close to the boundaries. To correct for this, we can instead estimate Cohen's *h* (Cohen, 1988). Cohen's *h* is a standardized difference in proportions, and can be calculated as

$$h = \varphi_1 - \varphi_2 \tag{3}$$

where φ represents the arcsine transformation of the estimated proportion, *p*, defined as

$$\varphi = 2 \arcsin \sqrt{p} \tag{4}$$

Cohen's *h* also has limitations. Note that in Equation 3 and Equation 4, the effect size is based only on the proportion of students achieving at a given performance level in each year. Neither sample size or the standard errors are included in the calculation. This can lead to misleading results when sample sizes are small, as a small change in the number of students in each performance level could lead to large percentage point changes. When calculating effect sizes for the DLM assessments, we use Cohen's *h* when the sample size is greater than 200 in both comparison years and Cohen's *d* when the sample size is less than 200. Both Cohen's *d* and Cohen's *h* can be interpreted on the same scale using the guidelines proposed by Cohen (1988) and expanded by Sawilowsky (2009). Effect sizes with a magnitude less 0.2 are *negligible*, between 0.2 and 0.5 are *small*, between 0.5 and 0.8 are *moderate*, and greater than 0.8 are *large*. By using Cohen's *h* as a default, we are able to evaluating changes in proportion without relying on a continuous measure. However, when the sample size is low (i.e., < 200), we can use Cohen's *d*, which can account for increased uncertainty that results from the smaller sample. Thus, when estimating changes in performance level distributions, we calculate the change in the percentage of students achieving at each

performance level, an effect size (either Cohen's *h* or Cohen's *d*, depending on sample size), and an effect size classification. In the following section, we discuss how to report these different outcomes.

# 4. Reporting Results

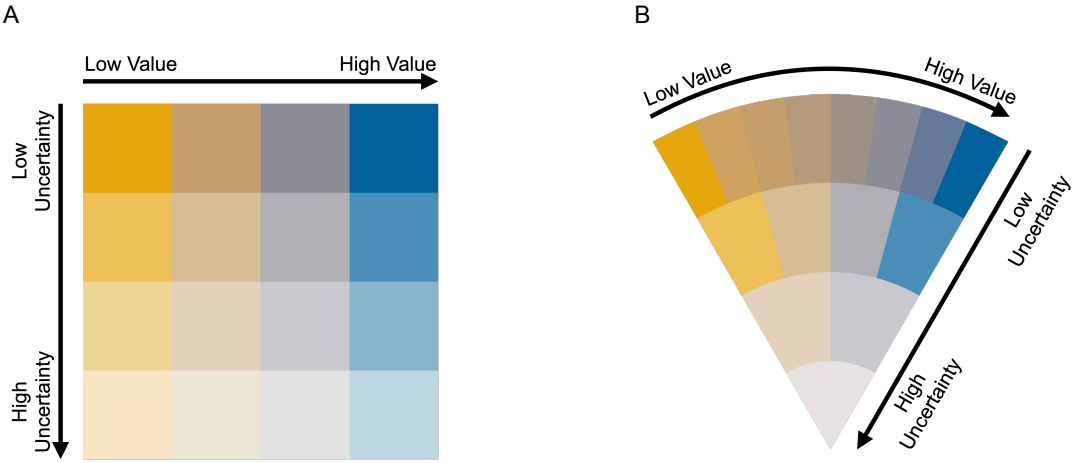
Reporting the results of the changes to performance level distributions requires us to provide as much information as possible without overwhelming the audience. For DLM assessments, there are four performance levels (*Emerging*, *Approaching the Target*, *At Target*, and *Advanced*) and one proficiency level (the percentage of students at the *At Target* or *Advanced* levels) for each grade, subject, and assessment model. This results in a total of 215 changes, effect sizes, and effect size classifications. To effectively communicate these outcomes, results are presented in three tables: Instructionally Embedded ELA and mathematics, Year-End ELA and mathematics, and science. The tables contain the estimated percentage point change for each performance or proficiency level, and table cells are shaded according to the effect size classification. This combination of values and shading allows readers to easily see the estimated change, as well as an indication of importance based on the shading. In the following sections, we describe the method and rationale for how color is applied to the tables and provide a demonstration of the method using the 2017–2018 to 2018–2019 comparison.

## 4.1. Table Shading Methodology

The table shading uses a value-suppressing uncertainty palette (VSUP), as described by Correll et al. (2018). VSUPs are bivariate color scales that normally show the raw value on one axis, and the uncertainty on the other. Figure 3 shows a traditional bivariate color palette compared to the VSUP. The VSUP uses a branching structure, such that there is greater differentiation of the colors when uncertainty is low. As uncertainty increases, there are fewer colors, with lower discrimination. In this way, VSUPs can discourage unintended conclusions when uncertainty is high.

**Figure 3**

*Comparison of Bivariate Color Palette and Value-Suppressing Color Palette*

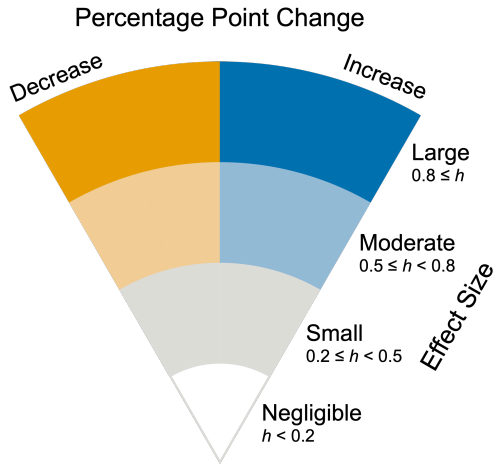



*Note*. Panel A: A standard bivariate color palette. Panel B: A value-suppressing uncertainty palette of the same color palette.

In the present analysis, rather than mapping the vertical axis to uncertainty, we use the effect size. That is, as the effect size gets closer to 0, the change has less practical significance, and therefore is represented with less differentiated colors. Additionally, for the changes in performance level distributions, we do not require the level of detail present in the full VSUP (Figure 3B). Rather, we adopt a reduced VSUP, as shown in Figure 4. In this version, the total number of colors has been reduced from the 15 in Figure 3B to only 6. For large and moderate effect sizes, decreases in the percentage of students at a given performance or proficiency level are shaded orange, and increases are shaded blue. Moderate effect sizes use a desaturated color to indicate that the effect is less strong. Finally, changes with small effect sizes are shaded grey, regardless of the direction of the change, and negligible effect sizes receive no shading.

**Figure 4**

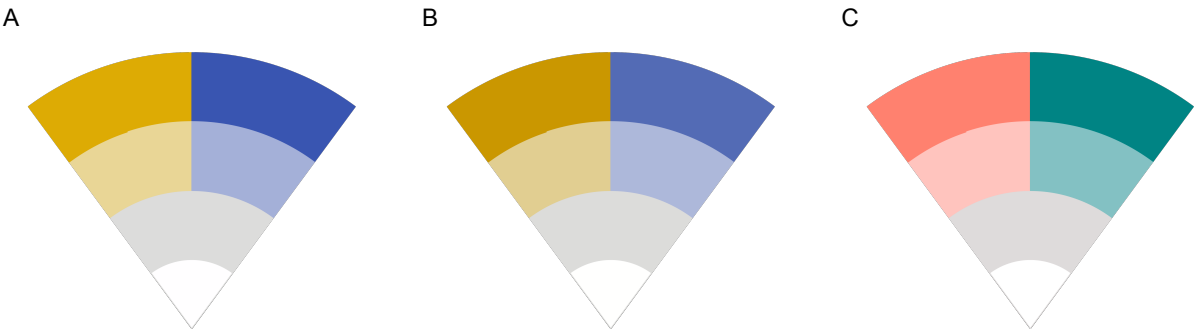
*Reduced Value-Suppressing Uncertainty Palette*



This specific color palette was chosen for two reasons. First, orange and blue are neutral colors that are not traditionally associated with positive or negative connotations (such as green or red). Neutrality in the color palette is beneficial because the same value of change could be seen as subjectively positive or negative depending on the context. For example, a 5-point increase in the percentage of students at a given performance level might be seen as a positive if the increase is occurring at the At Target level. Conversely, that same 5-point increase might be seen as a negative if the increase occurs at the Emerging level. Thus, we chose a palette that would not bias readers with unintended connotations. Second, the orange and blue color palette is accessible for individuals with color vision deficiency (i.e., color blindness). Figure 5 shows a simulation of what the palette in Figure 4 looks like under the three most common forms of color blindness: deuteranomaly, protanomaly, and tritanomaly (Simunovic, 2010). Under all three forms of color deficiency, the orange and blue color palette can be easily differentiated.

**Figure 5**

*Simulation of Selected Color Palette with Common Forms of Color Blindness*




*Note*. Panel A: Simulation of deuteranomaly, a form red-green color blindness. Panel B: Simulation of protanomaly, another form of red-green color blindness. Panel C: Simulation of tritanomaly, a form of blue-yellow color blindness.

In summary, the reduced-VSUP used for the performance level changes is able to communicate both the direction of the change (i.e., increasing or decreasing), as well as the practical importance, as defined by the effect size. Additionally, the chosen color palette is accessible and able to communicate this information without unintended connotations. In the next section, we apply this visualization method to observed data.

## 4.2. Example Reporting

Table 7, Table 8, and Table 9 present the example results for the Instructionally Embedded model, the Year-End model, and science, respectively, using 2017–2018 and 2018–2019 data. The numbers in the tables reflect the percentage point changes for each performance level from 2017–2018 to 2018–2019, adjusting the 2017–2018 sample using the propensity score matching described above. For example, the percentage of students in grade 3 ELA for the Instructionally Embedded model who achieved at the Emerging level decreased by 1.1 percentage points from 2017–2018 to 2018–2019 (Table 7). The cell shading in the tables reflects the importance of the percentage point changes for each performance level, as defined by the effect size of the change. For example, the percentage point change of −1.1 for grade 3 ELA at the Emerging level in the Instructionally Embedded model has no shading, which indicates the

percentage point change was negligible. To provide a second non-negligible example, the percentage point change for grade 10 mathematics at the Emerging level in the Year-End model was 13.0. This cell is shaded grey, which indicates a small effect size.

In total, the Emerging level for grade 10 mathematics in the Year-End model was the only non-negligible effect size across all models, subjects, and grades. Some changes that were large in magnitude were identified as negligible changes. For example, in grade 3 science, the percentage of students achieving at the Emerging performance level decreased by 22.3 percentage points in 2018–2019. However, the sample size for this grade was relatively small in 2017–2018 ($n$ = 145), as no states tested grade 3 for accountability purposed in that year. In this instance, the small sample size results in an effect size of −0.15, which is just short of the 0.2 magnitude that is required for a small effect size. Thus, overall, the results suggest that the performance level distributions were stable from 2017–2018 to 2018–2019, after adjusting for population differences.

**Table 7**

*Instructionally Embedded Performance Level Changes, by Grade, 2017–2018 to 2018–2019*

| Performance level | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|
| English language arts | | | | | | | | | |
| Emerging | −1.1 | 0.3 | 0.4 | 1.0 | −1.7 | 4.0 | −3.3 | — | −2.0 |
| Approaching the Target | 2.1 | −0.8 | −0.2 | 0.3 | 0.7 | −3.7 | −0.7 | — | 1.9 |
| At Target | −0.4 | −0.4 | 2.5 | −1.4 | 0.1 | 0.1 | 3.1 | — | −0.6 |
| Advanced | −0.6 | 0.9 | −2.7 | 0.1 | 0.9 | −0.4 | 0.9 | — | 0.6 |
| *At Target/Advanced* | *−1.0* | *0.5* | *−0.3* | *−1.3* | *1.0* | *−0.2* | *4.0* | — | *0.1* |
| Mathematics | | | | | | | | | |
| Emerging | 1.1 | 0.3 | 2.4 | −1.7 | −3.9 | −0.2 | −4.3 | −2.8 | −8.2 |
| Approaching the Target | 1.6 | −1.9 | −2.5 | 2.5 | 0.8 | −2.8 | 2.5 | 0.7 | 2.9 |
| At Target | −2.1 | 0.8 | 0.7 | −2.1 | 1.3 | 1.8 | 1.4 | 0.4 | 3.5 |
| Advanced | −0.6 | 0.8 | −0.5 | 1.4 | 1.8 | 1.2 | 0.4 | 1.7 | 1.9 |
| *At Target/Advanced* | *−2.7* | *1.6* | *0.1* | *−0.7* | *3.2* | *3.0* | *1.8* | *2.1* | *5.4* |

*Note*. English language arts is grade banded in grades 9–10 and 11–12.

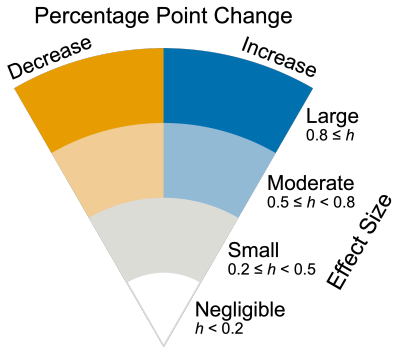


**Table 8**

*Year-End Performance Level Changes, by Grade, 2017–2018 to 2018–2019*

| Performance level | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|
| English language arts | | | | | | | | | |
| Emerging | 1.7 | 1.4 | 2.0 | 0.9 | −0.5 | −0.3 | 2.5 | 5.1 | 0.7 |
| Approaching the Target | 0.7 | 4.0 | 2.1 | 1.3 | 2.9 | 1.8 | 4.0 | −1.5 | 1.6 |
| At Target | −1.2 | −3.7 | −3.7 | −0.3 | 1.4 | −0.5 | −3.5 | −3.1 | −0.5 |
| Advanced | −1.2 | −1.6 | −0.4 | −1.9 | −3.8 | −1.0 | −3.1 | −0.4 | −1.8 |
| *At Target/Advanced* | *−2.4* | *−5.4* | *−4.1* | *−2.2* | *−2.4* | *−1.5* | *−6.5* | *−3.6* | *−2.3* |
| Mathematics | | | | | | | | | |
| Emerging | 1.4 | 2.6 | 2.7 | 0.6 | 1.4 | 1.3 | 5.2 | 13.0 | 0.1 |
| Approaching the Target | −0.6 | −1.5 | −0.4 | 0.9 | −0.5 | −0.1 | −1.4 | −7.3 | 2.2 |
| At Target | 0.2 | −0.4 | −1.3 | −0.4 | −0.3 | −1.0 | −2.3 | −5.5 | −2.2 |
| Advanced | −1.0 | −0.7 | −0.9 | −1.1 | −0.7 | −0.2 | −1.6 | −0.1 | −0.1 |
| *At Target/Advanced* | *−0.8* | *−1.1* | *−2.2* | *−1.5* | *−1.0* | *−1.3* | *−3.8* | *−5.7* | *−2.3* |

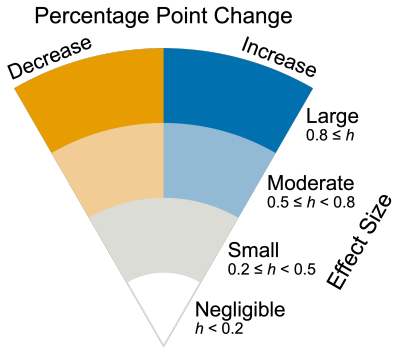
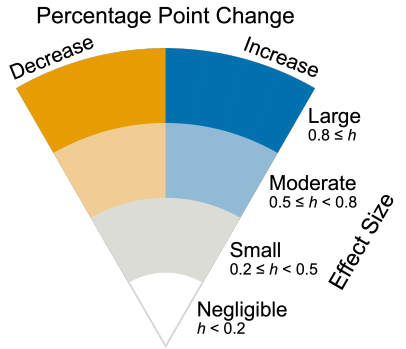


**Table 9**

*Science Performance Level Changes, by Grade or Course, 2017–2018 to 2018–2019*

| Performance level | 3 | 4 | 5 | 6 | 7 | 8 | 9–12 | Biology |
|---|---|---|---|---|---|---|---|---|
| Emerging | −22.3 | −0.9 | 2.4 | −7.0 | −12.4 | 1.5 | 0.7 | −1.1 |
| Approaching the Target | 15.8 | −0.1 | −1.7 | −2.2 | −1.6 | −0.1 | 0.7 | −0.7 |
| At Target | −1.3 | 0.2 | −0.8 | 6.3 | 9.5 | −1.6 | −1.1 | −0.9 |
| Advanced | 7.9 | 0.8 | 0.0 | 2.8 | 4.5 | 0.1 | −0.4 | 2.7 |
| *At Target/Advanced* | *6.5* | *1.0* | *−0.7* | *9.2* | *14.0* | *−1.4* | *−1.4* | *1.8* |



# 5. Discussion

In this report, we described and demonstrated a method for evaluating changes to performance level distributions when there has been a disruption to the population of students completing the assessment. This method is applicable to both acute disruptions (e.g., the COVID-19 pandemic), and long-term systematic changes (e.g., state compliance with ESSA 1% threshold). Specifically, we utilized propensity score models to resample students from previous administrations to more closely resemble the current demographic makeup of the population. This approach allows us to estimate what the performance level distributions would have looked like in prior years if the students' educational experiences were unchanged, but the population had demographic characteristics similar to the current population. Thus, any observed changes to the performance distributions would be a result of factors other than population changes.

To demonstrate this method, we applied a propensity score matching approach to data from the 2017–2018 and 2018–2019 administrations of the DLM alternate assessments. Results showed that the propensity score matching method proposed in this report was successful in reducing the differences between the populations from the two administration years. By accounting for population differences across the administrations, we were able to more effectively evaluate the changes in the performance level distributions as reflections of changes in performance rather than changes in population. Further, we presented a method for visualizing the changes that shows the magnitude of the changes while also allowing readers to quickly identify which changes were notable.

Throughout the report, we demonstrated our approach with data from the DLM assessments; however, these methods are can be applied to any operational assessment program. Although we selected a random forest model to use with the DLM assessments, we described a process for comparing and selecting from a set of possible propensity score models (Appendix A). Thus, following this approach, psychometricians in other assessment programs have the flexibility to select the propensity score model that is most appropriate for their data and student population. Additionally, the propensity score matching approach described in this report can be used for both acute and long-term systematic changes to the population of students participating in an assessment. Some changes to the student population have already happened (e.g., those resulting from COVID-19 and the "opt-out movement"), while other changes are ongoing (e.g., compliance with the ESSA 1% threshold for AA-AAS participation). Undoubtedly, there will be unforeseen disruptions in the future that cause additional changes to student population. The methods described in this report are not specific to any particular assessment or change in the student population. Therefore, these methods can be adapted for any situation, present or future, in which there was potentially a change to the student population across years, but an assessment program needs to evaluate changes in performance level distributions.

In any application, the propensity score matching method described in this report is constrained by the available data. For the DLM assessments, we used baseline demographic variables and responses to the First Contact survey as predictors in the propensity score model. However, there may be other important variables that were not included in this set of variables. If there were other factors that influenced assessment participation across years which are not included, the matched sampling may not fully capture all of the differences between the two administrations. Other applications of this work, and future research on these methods, should carefully consider potential variables to include in the propensity score model and evaluate the impact of leaving out important predictors on the effectiveness of matching algorithm.

In summary, this report demonstrated a framework for estimating changes in performance distributions across years when the student population has been disrupted, either intentionally or due to exigent circumstances. While this report specifically demonstrated the framework for the DLM alternative assessments, the methods can be applied to other assessment programs to facilitate cross-year comparisons that may otherwise be difficult or inappropriate. By accounting for population differences across years, we can make stronger claims that observed differences in performance level distributions are truly due to changes in performance, rather than a change to population of students completing the assessment.

# References

Agresti, A., & Coull, B. A. (1998). Approximate is better than ”exact” for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119–126. https://doi.org/10.1080/00031305.1998.10480550

Austin, P. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

Bennett, R. E. (2016). *Opt out: An examination of issues* (Research Report RR—16-13). Educational Testing Service. https://doi.org/10.1002/ets2.12101

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, *16*(2), 101–133. https://doi.org/10.1214/ss/1009213286

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

Chen, T., & Guestrin, C. (2016, August 13–17). *XGBoost: A scalable tree boosting system* [Paper presentation]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA. 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., & Li, Y. (2021). *xgboost: Extreme gradient boosting* [R package version 1.5.0.1]. https://github.com/dmlc/xgboost

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd). L. Erlbaum Associates.

Correll, M., Moritz, D., & Heer, J. (2018, April 21–26). *Value-suppressing uncertainty palettes* [Paper presentation], *Proceedings of the 2018 CHI conference on human factors in computing systems*. CHI ’18: CHI Conference on Human Factors in Computing Systems, Montreal, Quebec, Canada. 1–11. https://doi.org/10.1145/3173574.3174216

Cui, Z. (2020). Working with atypical samples. *Educational Measurement: Issues and Practice*, *39*(3), 19–21. https://doi.org/10.1111/emip.12360

de Rooij, M., & Weeda, W. (2020). Cross-validation: A method every psychologist should know. *Advances in Methods and Practices in Psychological Science*, *3*(2), 248–263. https://doi.org/10.1177/2515245919898466

DeMaris, A. (1995). A tutorial in logistic regression. *Journal of Marriage and the Family*, *57*(4), 956–968. https://doi.org/https://doi.org/10.2307/353415

Every Student Succeeds Act, 20 U.S.C § 6301 (2015). https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

Flach, P., Hernández-Orallo, J., & Ferri, C. (2011, June 28–July 2). *A coherent interpretation of AUC as a measure of aggregated classification performance* [Paper presentation]. Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA. 657–664. https://dl.acm.org/doi/10.5555/3104482.3104565

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://www.jstatsoft.org/v33/i01/

Harvey, W. (1960). *Least-squares analysis of data with unequal subclass members* (Technical Report ARS-20-8). USDA National Agricultural Library.

Kearns, M., & Valiant, L. (1994). Cryptographic limitations on learning Boolean formulae and finite automata. *Journal of the ACM*, *41*(1), 67–95. https://doi.org/https://doi.org/10.1145/174644.174647

Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review*, *39*(4), 261–283. https://doi.org/https://doi.org/10.1007/s10462-011-9272-4

Kubat, M. (2017). *An introduction to machine learning* (2nd ed.). Springer International Publishing AG.

Marland, J., Harrick, M., & Sireci, S. G. (2019). Student assessment opt out and the impact on value-added measures of teacher quality. *Educational and Psychological Measurement*, *80*(2), 365–388. https://doi.org/10.1177/0013164419860574

Nash, B., Clark, A. K., & Karvonen, M. (2016). *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. https://dynamiclearningmaps.org/sites/default/files/documents/publication/First_Contact_Census_2016.pdf

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*, 21. https://doi.org/https://doi.org/10.3389/fnbot.2013.00021

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, *96*(1), 3–14. https://doi.org/https://doi.org/10.1080/00220670209598786

Powell, M., Hull, D., & Beaujean, A. (2020). Propensity score matching for education data: Worked examples. *Journal of Experimental Education*, *88*(1), 145–164. https://doi.org/10.1080/00220973.2018.1541850

Sawilowsky, S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, *8*(2), 597–599. https://doi.org/10.22237/jmasm/1257035100

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, & B. Yu (Eds.), *Nonlinear estimation and classification: Lecture notes in statistics* (pp. 149–171). Springer.

Searle, S. R., M., S. F., & Milliken, G. A. (1980). Population marginal means in the linear model: An alternative to least squares means. *The American Statistician*, *34*(4), 216–221. https://doi.org/10.1080/00031305.1980.10483031

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Simon, R. (2007). Resampling strategies for model assessment and selection. In W. Dubitzky, M. Granzow, & D. Berrar (Eds.), *Fundamentals of data mining in denomics and proteomics* (pp. 173–186). Springer US. https://doi.org/10.1007/978-0-387-47509-7_8

Simunovic, M. P. (2010). Colour vision deficiency. *Eye*, *24*, 747–755. https://doi.org/10.1038/eye.2009.251

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, *24*(1), 12–18. https://doi.org/https://doi.org/10.11613/BM.2014.003

Stoltzfus, J. C. (2011). Logistic regression: A brief primer. *Academic Emergency Medicine*, *18*(10), 1099–1104. https://doi.org/https://doi.org/10.1111/j.1553-2712.2011.01185.x

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*. https://doi.org/10.1016/j.aci.2018.08.003

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodology)*, *58*(1), 267–288. https://doi.org/https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

XGBoost Developers. (2020). XGBoost documentation. https://xgboost.readthedocs.io/en/latest

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

# A. Comparison of Propensity Score Models

Propensity scores were estimated for each model, grade, and subject combination independently. This was done so that changes in one grade could be accounted for, even if the overall population was stable. In the propensity score model, the administration year (i.e., 2017–2018 or 2018–2019) was predicted from a wide variety of student demographic variables and First Contact survey responses. The included predictor variables are shown in Table 1. Because students do not have a receptive communication band assigned, a sum score scale was used.

It should be noted that these models are not intended to be causal or explanatory in any way. It would not be reasonable to claim a student's set of demographic covariates *caused* them to be assessed in a given year. Rather, the goal of these models is purely prediction (Yarkoni & Westfall, 2017). Regardless of the underlying causal processes, we aim predict which year a student was more likely to be assessed in. For the purpose of evaluating performance level changes across years, achieving equivalent samples across years is a higher priority than understanding the specific causal processes that resulted in non-equivalent samples (Shmueli, 2010). With this goal in mind, three methods were examined for estimating the propensity scores for each student. In addition to the random forest model described in Section 2.1, we also considered a logistic regression with LASSO (Least Absolute Shrinkage and Selection Operator) regularization (Friedman et al., 2010) and a boosted tree system (Chen & Guestrin, 2016; Chen et al., 2021). All of these methods are machine learning algorithms that can be used for variable selection to improve out of sample prediction.

For the logistic regression models with a LASSO regularization term used in this study, a binary response variable is modeled using a linear combination of predictor variables along with a logit link function (DeMaris, 1995; Peng et al., 2002; Sperandei, 2014). Logistic regression algorithms strive to find the weighted linear combination of predictor variables that best predict the response variable (Stoltzfus, 2011). The best weighted linear combination of predictor variables is determined by minimizing the sum of the loss function (i.e., the discrepancy between the true classification and the model-predicted classification) and the LASSO regularization term (i.e., the penalty for model complexity). More specifically, LASSO regularization constrains the coefficients for less influential predictor variables to zero in order to prioritize parsimonious models (Tibshirani, 1996).

Boosted tree systems also comprise a multitude of trees (XGBoost Developers, 2020). While random forests capitalize on the Law of Large Numbers to ensure improved performance (Breiman, 2001), boosted tree systems apply boosting to weak learners to achieve improved performance (Schapire, 2003). Weak learners for boosted tree systems are shallow trees that perform only slightly better than random guessing (Kearns & Valiant, 1994). Because it is often easier to identify weak learners than strong learners, boosting allows for many weak learners to be combined such that model performance is drastically increased compared to any individual weak learner (Schapire, 2003). Boosted tree systems add relatively shallow trees to the system one at a time, while using a gradient descent function to give weight to data points so that previous errors are corrected (Friedman, 2001; Natekin & Knoll, 2013). Adding trees based on the gradient descent function facilitates optimal learning of the structure of the trees in the system, which should improve model performance (XGBoost Developers, 2020). A weighted majority vote across all the trees in the system can then be used to determine the model classification.

Just like the random forest model, the LASSO logistic regression and boosted tree system also include

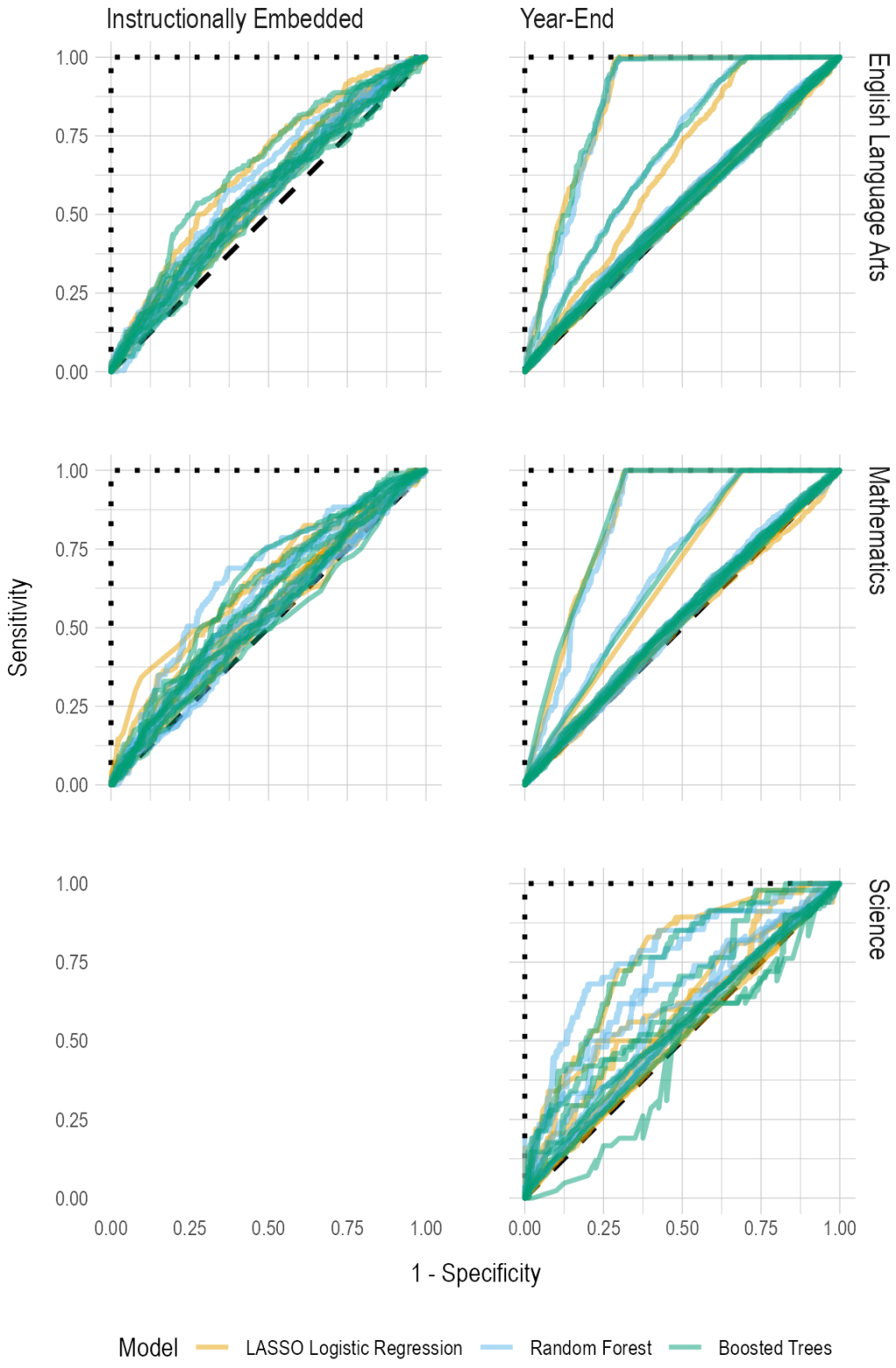
hyperparameters that cannot be estimated from the data. For the LASSO logistic regression, this is the penalty term that indicates the amount of regularization to perform. The boosted tree system includes the most hyperparameters. First, boosted trees include both of the hyperparameters that are included in the random forest (i.e., the number of predictors that are randomly sampled at each split when creating individual trees and the minimum number of data points in a node that are required for the node to split further). In addition to those two hyperparameters, the boosted tree system also includes parameters that control the total number of splits in a tree, the rate at which the boosting algorithm learns from iteration to iteration, the reduction in the loss function required to split further, and the amount of data exposed to the fitting routine. To estimate these model hyperparameters while also avoiding optimization bias, we used a nested resampling approach (see Figure 1). Model estimation then proceeds for each model as described for the random forest (Section 2.1).

Figure A.1 shows the ROC curves for the ELA and mathematics models and science models, respectively. In these figures, there is one curve for propensity score model and each grade or grade band (i.e., Instructionally Embedded ELA grades 9–10 and 11–12, and science grades 9–12). Curves that are pulled toward the upper left corner indicate higher predictive accuracy. The dotted line running through the diagonal of each panel represents random guessing. In general, we see that the curves for all three models are quite similar, with most models performing only slightly better than what would be expected by chance. This is not surprising, as we expect the population to be mostly stable across years. We expect changes to occur on the margins, as states enact policies to reach to the 1% ESSA threshold. In an ideal world where the population is completely stable, all the models would perform equal to chance (i.e., we cannot differentiate the population across years). The fact that we are able to predict at greater than chance, even marginally greater, indicates that there are slight changes to the population the models are able to detect. Thus, how much greater than chance we are able to predict can also serve as a rough measure of how much the population has shifted.

There are a few exceptions to the overall trends in Figure A.1. First, in Year-End ELA and mathematics, there are two sets of curves that are well above the others. These are grade 9 and grade 10. This is likely due to one state that changed their requirements for which high school grades are tested. This created a large shift in the observed demographics for these grades, making prediction easier for these grades. Second, there are few science curves that are further above the diagonal, although not to the extent of the grade 9 and 10 Year-End model curves. These curves are for grades 3, 6, and 7. These are the grades with the smallest sample size and were impacted by the addition of a new science state in 2018–2019. This state assesses science in all elementary and middle school grades and greatly increased the sample size in each those three grades (i.e., ~200 in 2017–2018 to ~700 in 2018–2019). Because these grades had such a large proportional increase in sample size, and because that increase almost exclusively came from one state, the models were able to more accurately predict administration year. Finally, there is one science curve that is noticeably below the diagonal, indicating performance less than chance. This curve is for the boosted tree model for the end-of-course Biology assessment, which has a very small sample size (2017–2018, $n$ = 169; 2018–2019, $n$ = 201). As discussed above, the boosted tree system includes a hyperparameter that controls how much of the available data is used in the estimation. Thus, this model uses an even smaller sample that what is available. Because the original sample sizes are already so small, it is likely that there is simply not enough data for this model to get a good estimation.

**Figure A.1**

*ROC Curves for Estimated Propensity Score Models*



We can quantify the performance of each model by calculating the total area under the ROC curve. An area of 1.0 indicates perfect prediction (i.e., the curve goes all the way up to the top left corner), and a value of .5 indicates chance guessing (i.e., the curve lies directly on the diagonal). Table A.1 shows the average accuracy (proportion of correct predictions), and Table A.2 shows the area under the ROC curve, across grades or grade bands. Overall, the random forest model tends to outperform both the LASSO logistic regression and boosted tree models. The median area under the ROC curve is higher for the random forest model in science and Year-End ELA and mathematics, and is comparable for Instructionally Embedded ELA and mathematics. Additionally, the average accuracy is higher for the random forest model for everything except Instructionally Embedded mathematics. Based on these results, the random forest was selected as the final propensity score model to be used for each grade and subject.

**Table A.1**

*Accuracy of Propensity Score Models, Across Grades*

| Subject | Logistic regression | Random forest | Boosted trees |
|---|---|---|---|
| Instructionally Embedded | | | |
| English language arts | .559 | .558 | .558 |
| Mathematics | .547 | .536 | .568 |
| Year-End | | | |
| English language arts | .513 | .522 | .518 |
| Mathematics | .510 | .522 | .520 |
| Science | .537 | .533 | .533 |

**Table A.2**

*Area Under the Receiver Operating Curves of Propensity Score Models, Across Grades*

| Subject | Logistic regression | Random forest | Boosted trees |
|---|---|---|---|
| Instructionally Embedded | | | |
| English language arts | .571 | .569 | .569 |
| Mathematics | .555 | .543 | .547 |
| Year-End | | | |
| English language arts | .518 | .530 | .523 |
| Mathematics | .518 | .533 | .529 |
| Science | .565 | .586 | .549 |