Putting Ducks in a Row: Methods for Empirical Alignment of Performance Standards

National Council of Measurement in Education
Vancouver, British Columbia

Katie Larsen McClarty
Daniel Murphy
Leslie Keng
Ahmet Turhan
Ye Tong

April 2012
Abstract

There is much focus at the state and national levels in graduating students that are prepared for college and careers. In order for students to be prepared at the end of their K-12 education, indicators are also needed along the way about whether students are on track. Using historical state data, nine different methods were used to align performance standards in mathematics grades 3-8 (i.e., OLS regression – forward, OLS regression – backward, logistic regression, quantile regression with 40th, 50th, and 60th percentile growth, equal percent impact data, vertical scale – equal intervals, and vertical scale – based on changes in test difficulty). The methods were evaluated for classification or prediction accuracy, number of correct classifications, pass rates, and types of errors. The effect of selecting different methods on standards alignment are discussed as well as potential implications for interpretations of student progress and school success.

Keywords: standard setting, vertical articulation, vertical alignment, vertical moderation, empirical alignment, evidence based standard setting
Putting Ducks in a Row: Methods for Empirical Alignment of Performance Standards

There is much attention at the state and national levels in graduating students that are prepared for college and careers. This notion of preparing students for learning beyond high school is also the focus of many school districts and high schools (Conley, 2010). Several sets of college and career readiness content standards have been developed (e.g., Standards for Success, American Diploma Project, College Board, ACT) to describe what students should know and be able to do in order to be prepared for postsecondary opportunities. Assessments have been developed to measure those skills, and standards have been established indicating college or career ready levels (Achieve & Pearson, 2009; ACT, 2007).

The current focus is on identifying levels of performance that indicate college readiness, are internationally benchmarked (e.g., compared with the Program for International Student Assessment), and will make students competitive in a global economy. Several methods have been suggested recently to specifically address setting college readiness performance standards (Haertel, Beimers, & Miles, 2012; O’Malley, Keng, & Miles, 2012) and internationally benchmarking standards (Phillips, 2012). Equally important, however, is setting performance standards at each prior level indicating students will eventually be on track for the college and career readiness goal. In order for students to be prepared at the end of their K-12 education, indicators are needed along the way about whether students are on track. Aligning the performance standards at prior grade levels can provide an “on track” or “early warning” indicator for students, teachers, parents, etc. Students who are not achieving appropriate progress can receive interventions to get back on track and graduate ready for a variety of postsecondary opportunities.
This type of performance standards alignment is being planned by the Race to the Top assessment consortia. The Partnership for Readiness for College and Careers (PARCC) application states that they will “establish cut scores for earlier grades that indicate whether students in earlier grades are on track to meet the readiness standard by the end of high school. This process will include a method to articulate the standards across grades and tests to ensure that performance expectations are coherent and provide a meaningful on-track signal” (PARCC, 2010, p. 63).

Alignment of performance standards is not a new concept. Lewis (2001) remarked that “the proficient cutscore set at one grade of a multi-grade assessment begs to be consistent with the message sent by the proficient cutscores at other grades” (p.1). He suggested several approaches for developing consistent standards including review of impact data, review of the performance level descriptors, review of grade-level judgments by participants at other grade levels, interpolation of the standards for some grade levels, and using standard errors to make adjustments.

Some of the methods for developing aligned performance standards can occur prior to tests being administered. These include developmental expectations for students, articulated content standards, articulated performance level descriptors, and including cross grade-level perspectives during the meetings. Other approaches, however, require the tests to have been administered. These include comparison to external reference points, a panel to review the results of other standard setting panels, smoothing impact data percentages, vertical scale information, and longitudinal analysis (Cizek, & Agger, 2012).

The most straightforward approach to developing “on track” inferences about student performance would be to follow groups of students from one grade to the next. Whereas this data
may have been challenging to collect in the past, the PARCC will collect two years of data prior to setting performance standards (i.e., field test and operational test). Because of this timeline, it would be possible to follow groups of students from one year to the next.

Based on that data collection design, the PARCC application suggests that they may use backward regression or student growth percentiles (Betebenner, 2009), which apply quantile regression, to inform the articulation process and set “on track” scores for lower grades (PARCC, 2010). These are two possible methods, but there are several methods that could be used to empirically align performance standards across grades. The purpose of this research is to evaluate different methods of empirically articulating performance standards when longitudinal cohort data are available.

**Data**

Eight consecutive years of historical state data from grades 3-8 mathematics were used for this study. The grades 3-8 statewide mathematics assessments are reported on a vertical scale. For this study, the first field-test will be referred to as Year 0, and each subsequent operational administrated will be referred to as Year 1 through Year 7. Years 0 and 1 were used to develop aligned performance standards in this study, and Years 2 through 7 were used to evaluate those standards. Table 1 shows the years of data included in the study, how the data were used, and the types of tests and grade levels included in the analyses from each year.

Table 1

<table>
<thead>
<tr>
<th>Years of Data and Uses in the Study</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Year 0</strong></td>
</tr>
<tr>
<td>Developed Standards using Multiple Cohorts</td>
</tr>
<tr>
<td>Field Test (Grades 3-7)</td>
</tr>
</tbody>
</table>
Study Design

Grade 8 served as the “anchor” grade from which the other standards were determined. Using multiple cohorts of students tracked from the field-test year (Year 0) to the first operational year (Year 1), the following methods were implemented to link and align performance across grades.

1. OLS regression – forward
2. OLS regression – backward
3. Logistic regression
4. Quantile regression – 40<sup>th</sup> percentile growth
5. Quantile regression – 50<sup>th</sup> percentile growth
6. Quantile regression – 60<sup>th</sup> percentile growth
7. Equal percent impact data
8. Vertical scale – based on equal intervals
9. Vertical scale – based on changes in test difficulty

The methods were then evaluated by following the single cohort of students from grade 3 in Year 2 to grade 8 in Year 7 to evaluate the accuracy of the performance standards as an “on track” indicator. The following metrics were used to evaluate the alignment methods:

- Classification or predication accuracy
- Number of correct classifications for a student over 5 years
- Pass rates
- Evaluation of types of errors
Developing Performance Standards

In order to develop the different sets of performance standards, student cohorts were tracked from field-test to first operational test. Although it would be ideal to have been able to follow the same cohort of students from third to eighth grade, this type of data is rarely available prior to setting performance standards. The field-test and first operational test data pattern more closely mirrors the data that will be available to the PARCC when setting their performance standards following the 2014-2015 test administration.

Each of the nine methods for empirically aligning performance standards will be described below. It is important to note, however, that all methods used the same “anchor” standard for grade 8 mathematics, which corresponded to a scale score of 1000. From that point, each of the nine methods was implemented to determine aligned performance standards for the lower grade levels.

1. OLS Regression – Forward

Starting with the performance standard set for grade 8 mathematics, a grade 7 “on track” performance standard was determined. For the OLS forward regression, grade 8 Year 1 scores were regressed on grade 7 Year 0 according to equation 1 to determine the grade 7 scale score that best predicted the grade 8 scale score of 1000.

\[
Gr8_{SS} = \beta_0 + \beta_1(Gr7_{SS})
\]  

(1)

Then the grade 7 Year 1 scores were regressed on grade 6 Year 0 scores for a different cohort of students. The grade 6 aligned performance standard could be calculated based on the grade 7 scale score derived in the previous step. This process was repeated for each pair of adjacent grades down to grade 3.
2. OLS Regression – Backward

Starting with the performance standard set for grade 8 mathematics, a grade 7 “on track” performance standard was determined. For the OLS backward regression, grade 7 Year 0 scores were regressed on grade 8 Year 1 scores according to equation 2. This method considered students that met the grade 8 standard and looks backward to where those students most likely were in grade 7.

\[ Gr7_{SS} = \beta_0 + \beta_1 (Gr8_{SS}) \]  

(2)

Then the grade 6 Year 0 scores were regressed on grade 7 Year 1 scores for a different cohort of students. The grade 6 aligned performance standard could be calculated based on the grade 7 scale score derived in the previous step. This process was repeated for each pair of adjacent grades down to grade 3.

3. Logistic Regression

Logistic regression was used to predict the probability of passing in grade 8 based on grade 7 scores. First students were dichotomized as passing or failing grade 8 based on scale scores where scores of 1000 or greater were classified as passing. Then the likelihood for students passing grade 8 Year 1 based on students grade 7 Year 0 scores were calculated according to equation 3.

\[ P(Gr8_{1000}) = \frac{1}{1 + e^{-z}} \text{ where } z = \beta_0 + \beta_1 (Gr7_{SS}) \]  

(3)

The grade 7 score with a 50% likelihood of passing grade 8 was used to set the aligned standard for grade 7. Then the grade 6 aligned standard was determined by calculating the probability of passing at that grade 7 standard based on the cohort of students that were in grade 7 in Year 1 and grade 6 in Year 0. This was repeated for all grades.
4-6. Quantile Regression

Starting with the performance standard set for grade 8 mathematics, three different grade 7 “on track” performance standards were determined using quantile regression methods. Quantile regression (Koenker, 2005) estimates a conditional linear quantile function,

\[ Q(\tau | X = x) = x' \beta(\tau) \]  

(4)

where \( Q(\tau) \) is the \( \tau \)th quantile of random variable \( Y \) and \( \beta(\tau) \) is the set of regression coefficients. The quantile regression procedure minimizes an asymmetric loss function for each \( \tau \) in a specified set \( T \subset (0,1) \), in particular for this study \( T = \{40, 50, 60\} \).

Cut scores for each grade were then established based on assumed growth at each of the three percentiles by chaining together predictions from adjacent cohorts. For example, the first step in creating the cut score system based on 40th percentile growth established the grade 7 cut score by estimating grade 7 to grade 8 growth. In particular, the established cut score was the grade 7 Year 0 score at \( \tau = .40 \) with the predicted score that was closest to a grade 8 Year 1 score of 1000. The next step established the grade 6 cut score as the grade 6 Year 0 score at \( \tau = .40 \) with the predicted score that was closest to the grade 7 Year 1 cut score. This process was repeated for each pair of adjacent grades down to grade three (i.e., grade 5 to grade 6, grade 4 to grade 5, grade 3 to grade 4). The same process was used to generate cut score systems based on 50th percentile and 60th percentile growth estimates.

7. Equal Percent Impact Data

The equal impact data method is similar to those proposed by Lewis (2001) and Cizek and Agger (2012) where the consistency of impact data is evaluated across grades. For this method, however, the consistency of impact data was based on the same group of students. To start, the percentage of students passing at grade 8 Year 1 was mapped to the score distribution
of those same students in grade 7 Year 0. The point on the grade 7 test mapping to the same percentage of students passing as grade 8 was used as the grade 7 standard. This was then repeated using grade 7 Year 1 students who were in grade 6 in Year 0 and so on.

8. Vertical Scale – Based on Equal Intervals

Based on the performance standard set for grade 8 (i.e., scale score 1000) and the performance standard set for the state in grade 3 (i.e., 600), performance standards for the other grades were set at equal intervals on the vertical scale (i.e. increasing 40 vertical scale score units each year). This approach is similar to one described by Ferrara, Johnson, and Chen (2004) where standards would be set at two end points (kindergarten and grade 3) and then linear interpolation would be used to derive the middle performance standards (grades 1 and 2) at equal intervals. A similar approach was implemented in South Carolina (Huynh, Meyer, & Barton, 2000). The difference here is that the equal intervals were based on the state’s vertical scale rather than linear interpolation. This approach assumes a linear growth trajectory for students in terms of mathematics progress from grades 3 to 8.

9. Vertical Scale – Based on Changes in Test Difficulty

For this method, the vertical scale was again used to align the performance standards that were set for the state from grade 3 (600 scale score points) to grade 8 (1000 scale score points). Rather than assuming linear growth as in method 8, however, this method used the state’s vertical linking constants to articulate the standards in theta units. If the state had a vertical linking constant between grades 7 and 8 of 0.460 and a grade 8 standard set at a theta value of 0.662, the aligned grade 7 standard would need to be set at a theta value of 0.202. This was calculated for each grade. Then the theta values were translated to scale scores for comparison with the other methods.
Aligned Performance Standard Results

As shown in Figure 1, using different methods for empirically aligning performance standards resulted in vastly different expectations for students. The performance standard for grade 3 ranges from a scale score of 606 using quantile regression with the expectation of 60\textsuperscript{th} percentile growth to a scale score of 841 using quantile regression with the expectation of 40\textsuperscript{th} percentile growth. The other methods fall in between.

*Figure 1*. Results of empirically aligned performance standards under nine different alignment methods.

For the most part, the trajectories of the performance standards are curvilinear with greater differences in the standards at the lower grade levels. This is to be expected because all methods started with grade 8 as the anchor point and then worked backwards using different cohorts of students that had two years worth of information. Differences between the methods
are most visible by third grade. The exception to the curvilinear pattern is the vertical scale based on equal intervals method which was designed to increase linearly.

**Evaluating Performance Standards Alignment**

The performance standards in Figure 1 were evaluated to determine whether they indicated that students achieving them were “on track” to reach the performance standards in the future. To evaluate the standards, data were used from a single cohort of students with complete mathematics data from grade 3 in operational Year 2 to grade 8 in Year 7. An analogous dataset could be constructed by following a group of students in the PARCC assessment system from spring 2016 (after the performance standards have been set) to spring 2021 when the third graders are in eighth grade.

Four different types of evaluation criteria were used. The first was classification or prediction accuracy, which is at the heart of the alignment interpretation. Do students who meet the standard (and are therefore interpreted as being on track) go on to meet the standard again the next year? The second criterion was the number of correct classifications for an individual student over the five year time period. To what extent are the performance standards providing a clear and consistent message for individual students? The third criterion was pass rates. Systems of performance standards could be extremely accurate if all students were classified as passing or all students classified as failing in every grade. Because of that, it was also important to evaluate the percentage of students passing at each grade level to see if the numbers were reasonable and appropriate. Finally, an evaluation of the types of errors was conducted. Some methods may be more likely to provide false positive indications (implying students are on track when they aren’t) whereas others may be more likely to provide false negatives (implying students are not on track when they are). Evaluation of these types of errors along with consideration of which
type of error is more tolerable for a state or consortia can aid in selection of a methodology for empirical standards alignment.

**Classification or Prediction Accuracy**

Students were coded as passing (1) or failing (0) at each grade based on the performance standard. Students that passed across all five years were put into an “always pass” group. For this group, the on track indicator they received in grade 3 was accurate in that they did pass at grade 4. That on track indicator at grade 4 was shown accurate by passing in grade 5 and so on through grade 8. Students that failed all five years were put into an “always fail” group. This group of students would have failed in grade 3, implying that they were not on track, and they would have continued to fail every grade through grade 8. These two types of patterns are shown in Figure 2. As can be seen, the percentage of students that always pass is larger than the percentage of students that always fail. Indeed, for some methods (i.e., OLS-forward, logistic regression, quantile – 50, quantile – 60, and equal percent) less than 1% of students are in the always fail category. The quantile – 60 method produces the highest percentage of always passing students as well as the highest percentage of consistent classifications overall.
Consistent Classifications all Five Years

Figure 2. The percentage of students consistently classified as on track/passing or not on track/failing across the five years.

**Number of Correct Classifications**

To evaluate the number of correct classifications for an individual student, students were coded as passing or failing at each grade based on the performance standard. Consistency in year-to-year pass/fail information was evaluated to determine the number of years in which a student’s “on track” information was accurate. For example, a student that passed in grade 3, passed in grade 4, but failed in grade 5 would have received an accurate “on track” indicator in grade 3, but an inaccurate indicator in grade 4. Students were evaluated to determine the number of correct “on track” indicators that they would have received under each of the different methods. This could range from 0 (where students flipped between passing and failing each year)
to 5 (where students were consistently passing or failing all five years). Figure 3 displays the results of this analysis.

The percentage of students that received five correct indicators mirrors the consistency information shown in Figure 2. Figure 3, however, also reflects students that were not consistent all five years. When looking at students that received four or five correct classifications, the quantile – 60 method is still the most accurate, but OLS – forward, logistic regression, and equal interval impact data are not too far behind. The least accurate methods were quantile – 40 and OLS – backward.

![Number of Correct "On Track" Indicators](image)

*Figure 3.* This provides the number of correct on track indicators a student received. Correct indicators include students indicated as on track that pass as well as students indicated as not on track that fail.
Pass Rates

As stated previously, indicators can be accurate if all students always pass or all students always fail, but that doesn’t provide much information to improve education for the students. Therefore, the percentage of students passing at each grade was also evaluated and is shown in Figure 4.

A few interesting patterns can be seen. First, both the OLS – backward and quantile – 40 methods show an increasing trend of pass rates. That is, fewer students pass in grade 3 than ultimately pass in grade 8, which could imply that students are learning more as they advance through the system. For OLS – forward, logistic regression, quantile – 60, and equal percent impact data, the opposite pattern is shown. More students pass in grade 3 (nearly 100% of students) than ultimately pass in grade 8, which could imply that students are regressing over the years. For the vertical scale based on equal intervals method, the percent passing is fairly consistent across grade levels, and for the quantile – 50 and vertical scale based on difficulty or linking constants, the pattern is less clear. In all cases, students were performing fairly well on the assessment, ranging from nearly a 70-100% pass rate across all grades and methods.
Figure 4. This provides the percentage of students that passed at each grade level under the aligned standards for each method.

Evaluation of Types of Errors

Finally, the types of errors produced by each method were evaluated. Some methods may have more false positives (indicating students are on track when they aren’t) whereas others may have more false negatives (indicating students are not on track when they are). The difference in the messages sent to students, parents, and teachers by the different types of errors is important as they may impact the types of interventions and services students receive.

Figure 5 shows the number of false positives at each grade level for each method. These are students that were deemed as on track (i.e. passed) in the current grade but then went on to fail the next grade. In almost all cases, the false positive rate is below 10% of the students. Quantile – 60, which had shown the highest accuracy rates overall, does show an increasing false
positive rate across grade levels. Conversely, quantile – 40 which had a lower accuracy rate overall reports fewer false positives as students move up in grades. The transition from grade 5 to 6 seemed to be an outlier in the false positive trend pattern for several of the methods.

**False Positives - "On Track" but Failed Next Grade**

![Graph showing false positives across grades for different methods](image)

*Figure 5.* This provides percentage of students at each grade who were deemed as on track (i.e., passed) but then went on to fail the next grade.

Figure 6 shows the percentage of students receiving false negative information for each method at each grade level. These are students that were deemed as not on track (i.e., failed) in the current grade but went on to pass in the subsequent grade. For many methods, the false negative rates were smaller than the false positive rates, and the false negative rates increased as the grade level increased. The exceptions were the OLS – backward and quantile – 40 methods. These two methods were more likely to indicate that students were not on track when they were successful in the next grade, and they had higher false negative rates in the lower grade levels.
False Negatives - Not "On Track" but Passed Next Grade

<table>
<thead>
<tr>
<th>Method</th>
<th>Grade 3</th>
<th>Grade 4</th>
<th>Grade 5</th>
<th>Grade 6</th>
<th>Grade 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS - forward</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS - backward</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quantile - 40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quantile - 50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quantile - 60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equal Percent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VS - equal interval</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VS - difficulty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 6.** This provides percentage of students at each grade who were deemed as not on track (i.e., failed) but then went on to pass the next grade.

The results of the false positive and false negative analysis make sense in relation to the pass rate analysis. For the OLS – backward and quantile – 40 methods, fewer students passed at the lower grade levels and more students received false negative information because more students went on to pass by grade 8. For the OLS – forward, logistic regression, and quantile – 60 methods, more students passed at the lower grade levels and more false positives were seen because fewer students pass in the eighth grade.

**Discussion**

According to the comparison of these nine methods, the quantile – 60 method was the most accurate. At first glance, it may seem strange that a method assuming all students make a greater than average amount of annual growth is most accurate. One reason why the results may
suggestion that students did make an above average amount of growth is because the data used to set the initial empirically aligned standards were based on a field test in Year 0 and an operational test in Year 1. It is likely that students were less motivated on the field test than under operational conditions, leading to suppressed performance in Year 0. If both Year 0 and Year 1 data had come from motivated sources, it may be that an expectation of average growth would be more reasonable.

Most of the methods showed a decreasing trend in percentage of students passing the test from grade 3 to 8. One interpretation could be that students are struggling more as the content requirements increase across grades. Another interpretation could be that schools are not preparing students as well for their future as they get older. Other methods (i.e., quantile – 40% and OLS – backward), however, could lead to the opposite interpretation—that the content is getting easier, students are becoming more proficient, and schools are preparing students well for the future. The possible interpretations of pass rate patterns and student progress should be considered, and a communications plan should be developed as the performance standards and student performance are released.

Finally, a state or consortia will need to evaluate the instances of false positives and false negatives. False negatives are likely to put more students in accelerated learning opportunities than may be necessary and result in lower ratings for a school or teacher. Additional learning opportunities may put a strain on educational resources but may also be helpful to students that are on the bubble. On the other hand, false positives are likely to put fewer students in accelerated learning opportunities and result in higher ratings for schools and teachers. However, they may also give students and parents a false sense of security, and they may not seek additional learning opportunities when they are needed.
The question about method accuracy, however, is an important one. A well aligned system may not show 100% accuracy. When the performance standards are aligned, a failing score should be seen as a warning indicator that a student is not on track and needs additional academic help. If the education system is reaching those students, it would be expected that false negatives would appear because students are being remediated and succeeding in the future. That same trend would not necessarily be expected in this study because of the use of historical data. Students did not receive pass or fail designations based on these study standards. Rather, their pass or fail designations and the educational responses were based on the actual standards in place by the state.

Many factors will need to be weighed in an effort to align performance standards. As states and the RTTT assessment consortia develop their next generation of student assessments, it is important that the information provided by the assessment results is accurate and informative. One of the interpretations to be made from the performance standard at each grade level is that students reaching the standard are on track to meet that same standard in the future. Empirical analyses can be used to validate that inference. The methods selected for the empirical analyses, however, can affect the resulting performance standards in important ways. This study provides valuable information that can be used moving forward in aligning performance standards across grades.
References


http://www.pearsonassessments.com/adpbriefingbook


