The Center for Assessment has been supporting PACE in producing “comparable annual determinations” since the 2014-2015 school year. We have developed and refined a methodology that has worked well within the PACE context and has allowed us, our technical advisory committee, and the U.S. Department of Education (USED) to be confident in the results. However, there were some issues and concerns raised about the achievement level reports produced for the 2017-2018 school year. This memo reviews the basics of standard setting and summarizes the standard setting methods used for PACE. We then discuss the changes made during the 2017-2018 standard setting process and the issues observed with the resulting achievement levels. We end with recommendations for moving forward.

Standard Setting Methodology

Determining the achievement levels on PACE (e.g., proficient, advanced) or any other test involves matching achievement (sometimes called performance) level descriptors with scores on a test or other distribution. Achievement level descriptors (ALDs) are the foundation of this process. There are many classes of methods used to connect ALDs with score distributions. Some methods focus on having panelists review and evaluate the difficulty of individual test items, while others have panelists review collections of student work or portfolios of scores. Still others rely on capturing expert judgments of examinees. We have been using an examinee-centered method called contrasting groups. This standard setting method involves using judgments from panelists (i.e. teachers) about the overall achievement of the examinees based on extensive knowledge of the examinee. To implement this method for the PACE pilot, we ask teachers at the end of the school year to make judgments about which achievement level best describes each of their students. This process relies heavily on a common understanding and interpretation of the achievement level descriptors (ALDs). The subject and grade specific ALDs are entered into an online survey where teachers can easily read the descriptions and match their students to the appropriate achievement level. The teacher judgment data is one component of the standard setting approach. These judgments must then be matched against the scores on competency-based assessments the students have completed throughout the year. One expects the students rated highest by their teachers to have the highest competency scores and the students rated lowest to have the lowest scores. That is generally true, but as one would expect, there is a fair amount of overlap.
The contrasting groups standard setting methodology then involves comparing the PACE scores with student placements into achievement levels in order to determine cut scores that would accurately classify the highest percentage of students into achievement levels.

Logistic regression is used to determine the point in the score distribution where examinees have a 50% chance of being classified in the next performance level or above (e.g., the probability that a student is Level 3 or above is 50% at score X). A logistic regression analysis is run separately for each cut point—Level 2, Level 3, and Level 4—in each district, content area, and grade level. The figure below illustrates a hypothetical case for how logistic regression is used to estimate a cutscore between two distributions. Unfortunately, the real world is rarely as clean is the picture below, but the method has worked well through the years.

The participating PACE districts all rely on unique competencies and fairly unique assessments except for the PACE common task and perhaps reused PACE common tasks to make local competency determinations for students. Therefore, the logistic regression analyses are computed for every district/grade/subject combination.

**Issues with the 2017-2018 PACE Standard Setting**

Several new school districts and parts of school districts were included in PACE accountability system during the 2017-2018 school year. The assumptions underlying most statistical methods are challenged when dealing with small sample sizes. This challenge was significantly compounded by teacher judgment data that had considerably less variance than in previous years.
The PACE achievement levels have always been more constrained than with tests like Smarter Balanced or NH SAS that are designed to spread the distribution of students, but we observed noticeably more spread in judgments in previous years. Some of the most egregious cases this year involved classes with all 20 or so students receiving the same rating (e.g., “3”). We are pretty sure this cannot be accurate. This example is an extreme case, but the combination of generally more constrained distributions and smaller sample sizes made us less confident in the method we had been using since 2014-2015.

This played out by having too many cases where the logistic regression was unable to provide any estimate of a particular cutscore. In general, the cutscore between levels 2 and 3 was able to be estimated for almost all grade/subject/district combinations, but there were a substantial number of cases where the cutscores between levels 1 and 2 or between 3 and 4 could not be estimated. We are required to report student achievement in the same way (or as similar as possible) as the statewide assessment and in accordance with the ESSA technical requirements, which means that we need to produce three cutscores to yield four achievement levels. We observed very limited instances of these issues in the past and when we did, we “manually fit” a cutscore by, for example, taking half of the distance between the 2/3 cutscore and the highest possible score (e.g., 4.0) to produce a 3/4 cutscore. We were not always comfortable doing this “manual fitting” even with a very limited number of cases (e.g., why is halfway the right place to put the missing cutscore compared with some other distance?), but we became very concerned when it looked like we might have to do this for many of the smaller schools.

**Modified Standard Setting Approach for 2017-2018**

Given these issues, we worked hard and fast to invent a new approach for establishing cutscores for the 2017-2018 school year. The small sample sizes were our biggest challenge so we designed an approach to ameliorate that concern. The modified approach involved standardizing all district/grade/subject scores such that the mean score for each district/grade/subject combination was 0 and the standard deviation was 1. This allowed us to pool data from all districts having that particular subject/grade combination (e.g., 7th grade math). These standardized scores (z-scores) allowed us to estimate all three cutscores for every grade/subject combination while maintaining the relative rigor of the grading practices in each district intact. We then translated the z-scores back to the metrics used in each district (e.g., 1-4 scale, 1-100 scale) for reporting purposes. We examined the results of this new method compared to the original method, but we only did make these comparisons in the aggregate (e.g., 7th grade math for all PACE districts) and, on average, things looked comparable plus we were happy that we did not have to do any “manual fitting.”

**What Went Wrong?**

While the results appeared correct (again, on average), a call from SAU 39 at the end of November pointed out things did not work as well as it appeared on average. In hindsight, we should have done a much more thorough quality control process where we examined specific district/grade/subject results for a strategic sample of such combinations. To be fair we were on a
very compressed timeline to turn the cutscores over to the DOE data contractor, but we should have insisted on additional time to do these checks. SAU 39 reported significant declines (e.g., 30-40 percentage points) in the proportion of students scoring proficient and above for the same cohort of students (e.g., 4th grade in 2017 to 5th grade in 2018). Hearing this caused us to dig into the data to try to determine the extent of the issue.

We first examined the results from typically high and low scoring districts and began to suspect that the pooling procedure generally pulled the high scorers down and boosted the lower scoring districts by lessening true differences in achievement across districts. We communicated these concerns to the NH DOE on December 21st, but we indicated that we needed additional time to investigate the issue before doing something as drastic as replacing existing score reports with new reports. We continued to investigate the results and now feel quite sure that our modified standard setting method produced inaccurate results.

**Recommendations for Reporting the 2017-2018 Cutscores**

This may sound a little like a “rock and hard place” conundrum in that we are fairly confident that the modified method produced inaccurate cutscores for many districts even though it allowed us to estimate all cutscores for all districts. On the other hand, we suspect the original method produced more accurate cutscores but only when the statistical model was able to provide a reliable estimate. Therefore, we recommend the following steps:

1. Immediately flag or otherwise signify on the public release of the PACE results that the posted information must be considered tentative while the NH DOE investigates concerns with the accuracy of the posted scores.
2. Use the original model to produce cutscores for all district/grade/subject combinations even when manual adjustments are required. We have developed a set of established rules for making such adjustments as follows:
   a. Adjustment to the Highest Obtainable Scale Score (HOSS): When the cut score falls above of the obtainable competency score range, the cut score for Level 4 will be adjusted to the highest obtainable scale score.
   b. Adjustment to Midpoint: When the cut score cannot be estimated and falls between two estimated cut scores, the cut score will be determined to be the midpoint between the two estimated cut scores or between an estimated cutscore and the HOSS or LOSS (lowest obtainable scale score).
   c. Equipercentile Adjustment: When there is at least one missing cut scores on either side of the flagged cut score (e.g., Level 2 or Level 4 cuts), an equipercentile equating procedure is used to estimate the cut score that would closely replicate the distributions of achievement across the performance levels in the same district and subject for the other grade levels with unadjusted cut scores. In the few cases where there are no other grade levels with unadjusted cut scores, the same grade
level will be used in the other content areas to approximate the distribution of achievement.

3. Once the full set of original and modified cutscores are in hand, create a side-by-side comparison of original and modified cutscores for every district/grade/subject combination. Having two sets of scores in and of themselves is no different than the “two watch problem” (when you have one watch you know the time, but when you have two, you’re never quite sure). Therefore, we need criteria by which we can evaluate the efficacy of the two sets of cutscores. We propose using the following three types of comparisons to help shed light on the efficacy of the two methods. While such investigations rarely reveal an absolute conclusion, we are confident that the body of evidence will point us in one direction or another.
   a. Student longitudinal data that allows us to examine how the students in a given grade in 2018 performed in the previous grade (same subject) for the previous year (2017).
   b. Cohort-to-cohort data that allows us to compare the performance of students in a given grade/subject in 2018 to the same grade/subject in that district for the previous year (2017). This is generally not as robust as longitudinal comparisons, but it provides another way to view the results.
   c. Comparison of the PACE achievement level designations in a given grade span (e.g., elementary) to the NH SAS results for the same grade span.

4. Once the analyses are completed and a decision made, we recommend either removing the “flag” for the currently-reported scores or releasing new scores.

**Improving our Methods Going Forward**

We have and continue to consider a set of comprehensive improvements to the process going forward, but we will save these recommendations for a subsequent memo.