Recommendations for Improving PACE Standard Setting

Scott Marion and Carla Evans

National Center for the Improvement of Educational Assessment

January 27, 2019

We documented the concerns with the 2017-2018 achievement levels in a previous memo and analytic report. We have and continue to consider a set of comprehensive improvements to the achievement level determination process for 2019 and beyond including the following:

1. Improving training and communication for the teacher judgment process,
2. Creating summary ALDs,
3. Include the use of “impact data,”
4. Use of “equipercenitle adjustments” if necessary,
5. Testing additional methods using data from 2017 and 2018, and
6. Improving the data flow and quality control/quality assurance (QC/QA) processes.

We expand on each of these steps below.

1. Improving the training and communication associated with the teacher judgment process.
   As PACE has expanded, we question whether we have kept up with the training approaches necessary to ensure that teachers understand the importance of the teacher judgments and how they should produce these judgments. To this end, we will produce narrated slide presentations and would like cooperation from the district leads in requiring ALL teachers to view these training materials. The narrated slide deck will include many of the directions that reflect best practice from the past and will include many of the new points discussed below.

2. Creating “summary” achievement level descriptors (ALDs).
   The current PACE ALDs were based on the Smarter Balanced ALDs, which are very similar to the NH SAS ALDs. These ALDs are referred to as “range ALDs” and are used to describe the expected performance for students at a given level (e.g., proficient) for each of the many standards in a specific content area. It is critical that participants adhere closely to the descriptions of performance in ALDs when making the judgments about each and every student. We are concerned that the range ALDs make this task too cognitively challenging for participants. Therefore, we, along content experts, will draft summary ALDs designed to capture the essence of prototypical performance at each specific achievement level. Our goal is to anchor these summary ALDs with student work samples to help make ALDs more concrete. These summary ALDs will be used as the basis of teacher judgments, but not until 2020.
3. **Including “impact data” in the teacher judgment process.**
   Standard setting studies often include “impact data” so panelists can get a sense of how the recommended cut scores will translate to aggregate performance (e.g., percentage of students at each performance level). This is not possible to do with the teacher judgment method, but we suggest producing average distributions of students in each level based on the most recent three years of state test scores at the specific grade span. These impact data are not meant to constrain teachers’ judgments, but to help inform them of the relative performance at the school.

4. **Use of “equipercentile adjustments” if necessary.**
   We suspect, unfortunately, even following steps 1-3, that there will be cases when teacher judgment results are highly unrealistic (e.g., all students are marked as Level 3) or completely unrelated to students’ past performance. We will develop guidelines for what constitutes “unrealistic” and once such guidelines are agreed upon, we would apply an “equipercentile adjustment” to unrealistic teacher judgments. What would this look like? We would take a multi-year average of state test scores for the given grade span and subject area and apply the proportion of students in each level to the competency distribution to yield an equivalent proportion of students in each level on the competency score distribution. This is a bit of a heavy-handed approach that we would rather not apply, but will be forced to in rare cases of unrealistic results.

5. **Testing additional methods using data from 2017 and 2018.**
   We employed the “pooled method” this year in an effort to rectify concerns with the method we had been using since 2015. We have several ideas for continuing to improve the processes we use and figure out the most accurate and consistent standard setting process possible. This time, however, we will test potential new methods against existing data (2017 and 2018) and evaluate the results much in the same way that we compared the pooled and original standard setting methods. This way, we will have a transparently approved process prior to having to implement it in 2019.

6. **Improving the data flow and quality control/quality assurance (QC/QA) processes.**
   One of the major limitations in the current process is that we do not have a strong enough set of QC/QA protocols in place to catch the type of concerns that arose this year. In addition to improving the data flow processes, we will conduct analyses similar to the comparative analyses used to evaluate the efficacy of the pooled and original standard setting methods. Doing so will allow us to significantly reduce the chances of publishing results that raise the kinds of concerns experienced this year.