Independent Review of Analyses Conducted by Smarter Balanced

March 19, 2018

Joseph A Martineau
Senior Associate

Center for Assessment

Smarter Balanced conducted technical analyses to investigate the possibility of technical issues with the 2017 assessments based on perceived differences between the generally positive cohort-to-cohort gains from 2015 to 2016 and the generally flat cohort to cohort change scores from 2016 to 2017. There are in general three possible explanations for the differences described above:

1. The differences are attributable to *true differences* in student achievement.
2. The differences are attributable to *construct-irrelevant* factors.
3. The differences are attributable to *construct-relevant* (but unmodeled) factors.

Of these three potential explanations, (2) and (3) would be problematic, with (2) being the most problematic.

I have reviewed results and associated conclusions from the analyses conducted by Smarter Balanced within the framework of the three potential explanations. Several research questions were posed to evaluate hypotheses relating to potentially problematic explanations. Overall, I concur with Smarter Balanced's conclusion that the analyses did not point toward any procedural or technical errors.

In addition to reviewing Smarter Balanced's analyses and conclusions, I also provided suggestions for refining the current analyses and some guidance on methodology which Smarter Balanced may want to investigate as part of continuous improvement efforts.

I find the proposal presented in the report to directly test the hypotheses of students becoming more familiar with surface feature of novel items to be a compelling potential refinement of the analyses conducted in this report. The report also mentions additional potential avenues for research that may be productive. These additional proposed analyses represent thoughtful and appropriate responses to the fact that the analyses conducted do not point toward any problematic mechanism that could have produced the observed differences between changes in scores from 2015 to 2016 versus from 2016 to 2017. The analyses proposed should not be hurried for two reasons: (1) they are complex and require thoughtful design of the analyses, (2) the differences observed are very small in term of effect sizes, and (3) there is no compelling evidence as of yet that points to any problematic mechanism.

Review of Each Question, Associated Analyses, and Associated Conclusions

My review of each research question addressed is provided below, along with my recommendations, where appropriate, to refine the analyses conducted as a part of continuous improvement efforts.

How did students perform in 2017 compared to 2016?

I concur with the findings in the report. I also propose an alternate approach and point out that it would be helpful to more closely examine differences between jurisdiction, grades, and differences between 2015 and 2016 and 2016 and 2017.

Were there fewer test questions available?

I concur with Smarter Balanced's finding and the straightforward conclusions.

Did students receive more difficult test questions in 2017 compared to previous years?

I largely concur with the findings and conclusions. Although I do not anticipate it changing the conclusions, I would prefer for differences in item pools to be represented as the total available pool in 2016 and the total available pool in 2017.

Was the assessment differentially precise in 2016 and 2017?

I concur with the results and conclusions presented by Smarter Balanced.

Did the newly added test questions impact test results?

I concur that there are systematic differences between CAT-delivered items from the 2016 pool and newly-available CAT-delivered items in the 2017 pool, and performance task (PT) based items. I also concur that these differences between the three sets of items sets were very small. Finally, I concur that the analyses did not suggest a mechanism by which those differences could have arisen that is consistent with construct-irrelevant factors or construct relevant (but unmodeled) factors.

The differences are very small, and the analyses did not point to a problematic mechanism that could have produced the observed effects. Still, I assert that this line of investigation is the most promising for refining the work presented in this report to identify and evaluate any new potential hypothesized mechanisms. Consistent with this assertion, I recommend instituting refined analyses of this type as part of continuous improvement efforts. Specifically, I recommend keeping track of what item types are familiar vs. novel for each year of test administration, when items of each type (novel, familiar) are calibrated, what that would mean for items being less or more difficult than expected in each year of administration, what that would mean could be expected of item residuals in each year of administration, and finally, what that would mean could be expected of student scores in each year of administration.

Did students spend less time taking the test?

I concur with Smarter Balanced's analyses, results, and conclusions based on the structure used for the analysis. However, the analysis would be improved if time spent on operational items could be separated from time spent answering embedded field test items.

Did students take the test earlier in the school year?

I concur with Smarter Balanced's conclusions based on the analysis conducted. While I do not anticipate it would changes the conclusions, I would prefer to have jurisdiction-to-jurisdiction differences in test start date and school calendar taken into account.

Were the student demographics different?

I concur with Smarter Balanced's results, and conclusions.