Memo on ACT Sections and EOC Exams

Zack De, Ohio Department of Education

June 14, 2019

Background and Research Questions

Ohio first began requiring the vast majority of students take a college entrance exam (the SAT or ACT) in their junior year starting in spring of 2017, with the first cohort of students graduating in 2018. This memo details an initial look at the relationships between proficiency on Ohio's End-of-Course (EOC) tests in English Language Arts II, Geometry, and Integrated Math II exams and the ACT sub-scores. In particular, we seek to understand the following:

1) To what extent does performance on Ohio's EOC exams predict performance on the ACT?

2) How does the cut-score for proficiency on Ohio's EOC exams relate to the Ohio Department of Higher Education's remediation-free scores on the ACT (18 on English, 22 on Reading, 22 on Math)?

Methods

In order to answer the above questions, I built linear regression models and Receiver Operator Characteristic (ROC) curves for the following set of predictors and output variables (see below). Data come from Ohio's 2018 graduation cohort and include both students who did graduate and those that did not. Please note that there are significantly fewer students who take the Integrated Math II EOC assessment than the Geometry EOC assessment; however, there is still enough data in order to perform an analysis (more than 7,500 students).

Predictor	Output Variable	N
English Language Arts II EOC Exam	ACT English	109,491
English Language Arts II EOC Exam	ACT Reading	109,382
Geometry EOC Exam	ACT Math	73,067
Integrated Math II EOC Exam	ACT Math	7,572

Linear regression is used to show the total relationship between one variable (an EOC exam) to another (ACT sub-score) in a linear way. ROC curves, on the other hand, are built to show the tradeoff in a binary classification system between the true positive rate the false positive rate (more explained below).

Metrics, explained

We can further look at metrics for each particular point along a ROC curve. Below are just the ones I have included in this memo:

- True positive rate (true positives over real positives)
- False positive rate (false positives over real negatives)
- Accuracy
- Informedness

True positives are students who score at or above the EOC threshold and do, in fact, score remediation-free on the ACT while **false positives** are students who score at or above the EOC threshold, but **do not** score remediation-free on the ACT. **False positives** are of particular concern since these are students who would be identified as not needing extra support (but potentially needing it.)

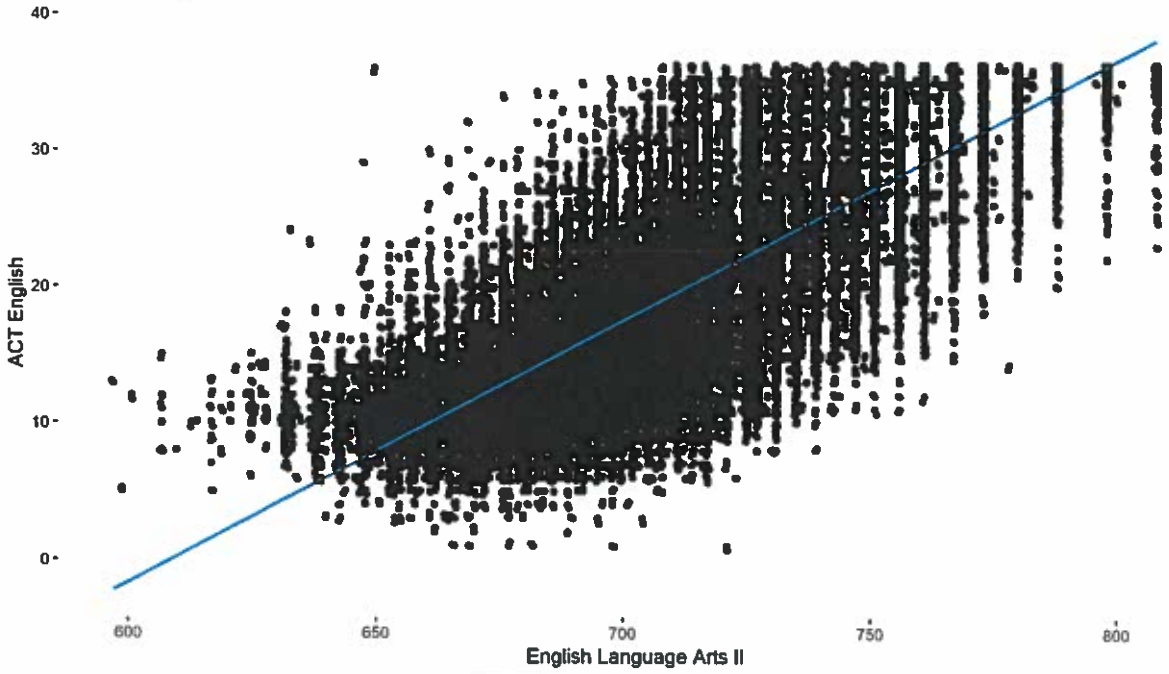
Additionally, we can also examine the related, but not identical metrics of *accuracy* and *informedness* for each scale score. **Accuracy** ranges from 0 to 1 and is the percentage of all predictions that were correct (true positives and true negatives over all real positives and all real negatives). **Informedness**, however, is the probability of an informed decision; values range from -1 to +1, with +1 representing the most correct and perfectly informed model, -1 representing the least correct, but most perfectly informed model, and 0 representing a minimally informed model. As a hypothetical example: if a cut-point were set at the very lowest scale score possible, every student would be predicted as scoring remediation-free on the ACT by virtue of taking the end-of-course exam. This would result in a model that has some degree of accuracy because many students would be remediation-free; however, it would be minimally informed because it does not discriminate between EOC test takers. As a result, one should consider both accuracy and informedness when evaluating different models.

Results, Linear Regression

Results show that for each analysis, about 55% of the variance is explained simply by performance on a related EOC test. Graphs have had jitter (small amounts of random noise) added to visualize the points.

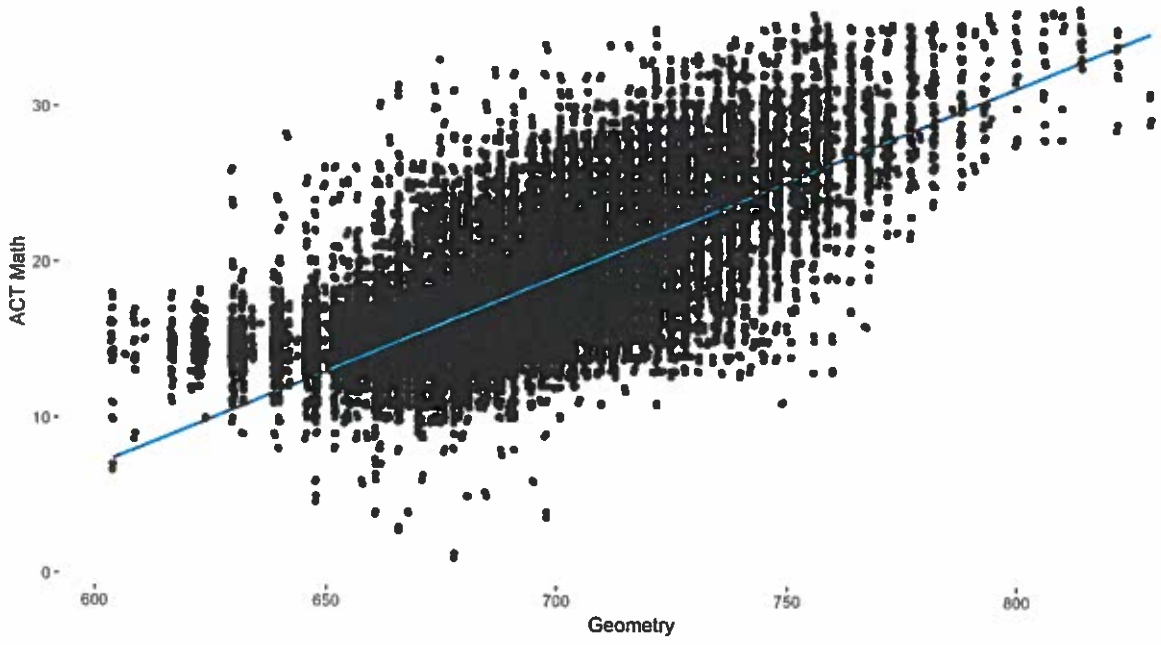
English Language Arts II vs ACT English

R-squared = 0.591



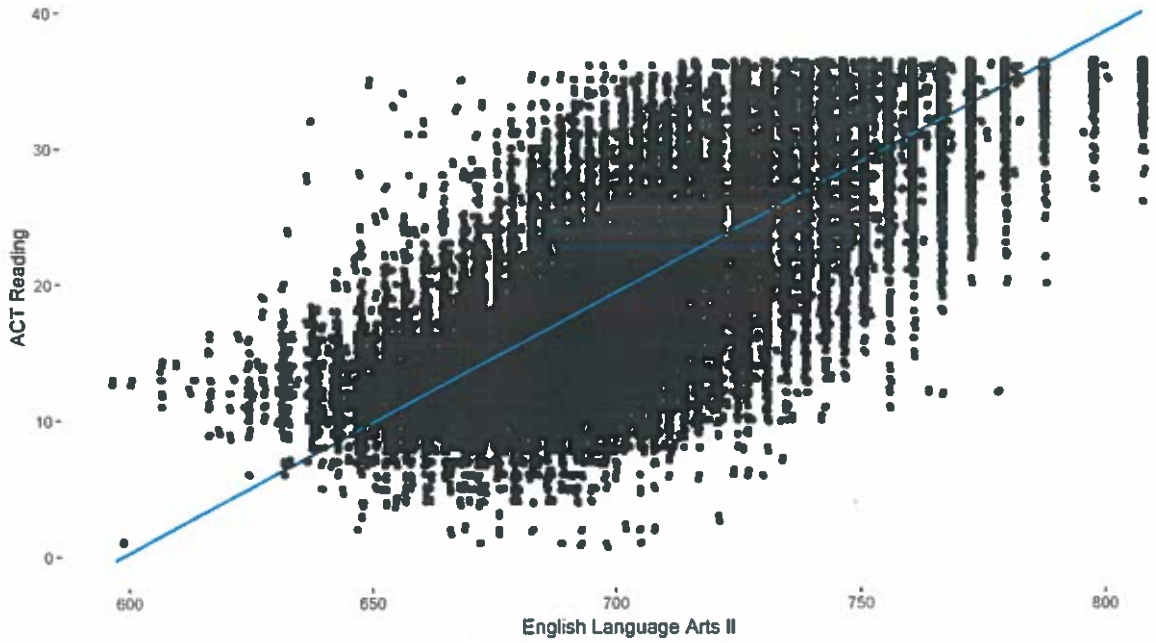
Geometry vs ACT Math

R-squared = 0.551



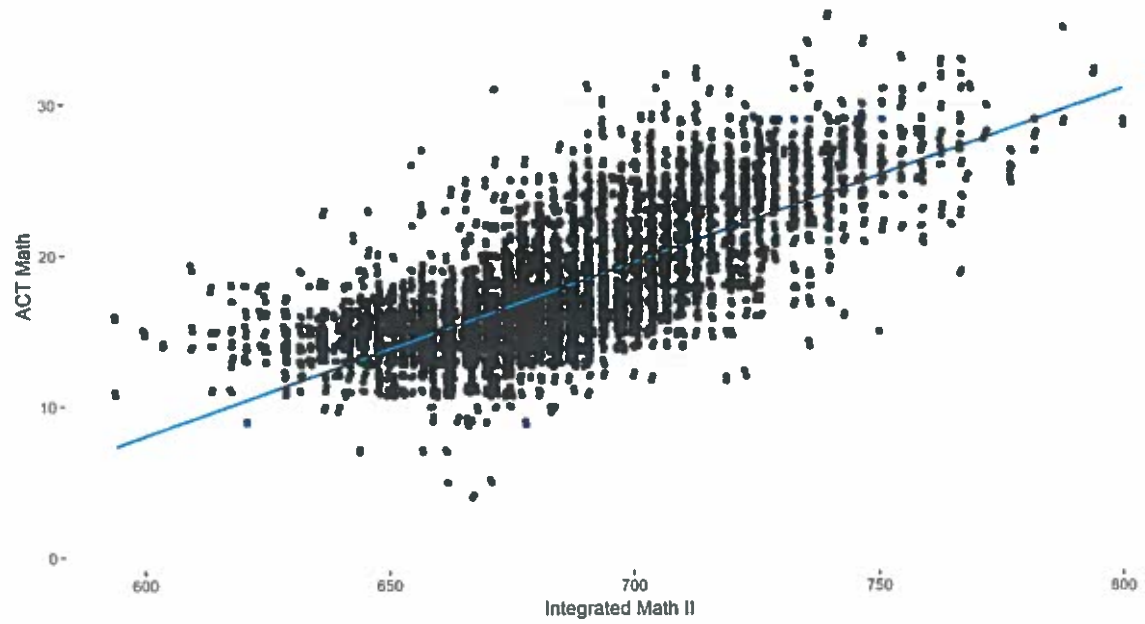
English Language Arts II vs ACT Reading

R-squared = 0.586



Integrated Math II vs ACT Math

R-squared = 0.548



Results, AUC

The Area Under the Curve (AUC) for the ROC curves are built specifically with the remediation-free ACT scores in mind. The AUC listed below measures the "goodness" of the entire ROC curve, rather than at any specific point. AUC numbers range from 0.5 (no better than chance) to 1 (perfect prediction), with .9 generally considered excellent.

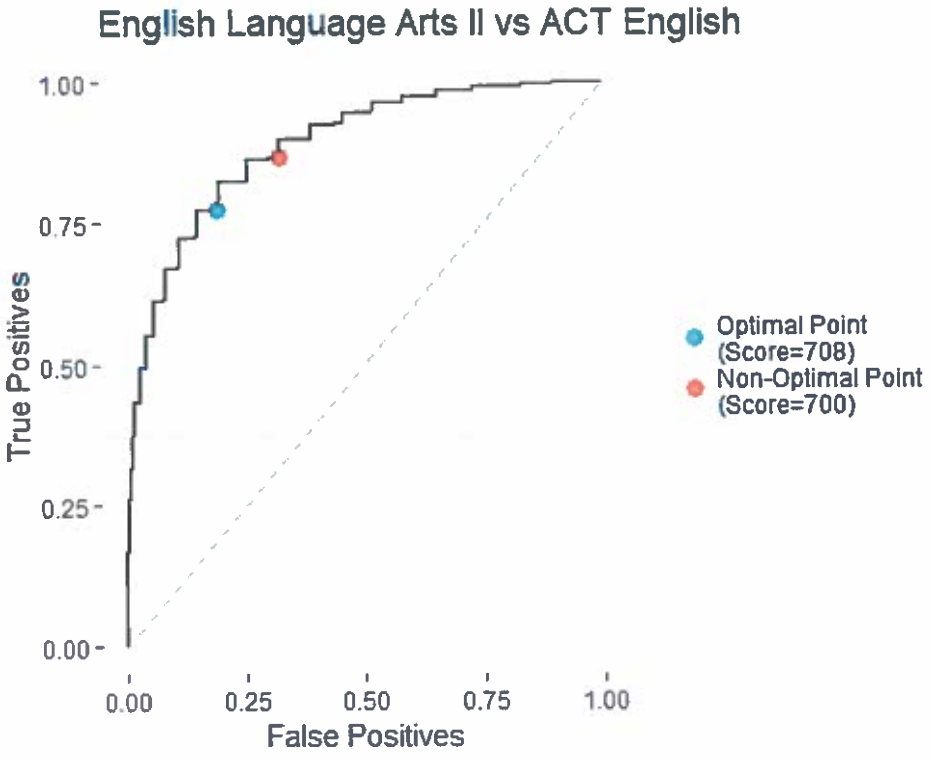
Predictor	Output Variable	AUC
English Language Arts II EOC Exam	ACT English	0.8866161
English Language Arts II EOC Exam	ACT Reading	0.8947938
Geometry EOC Exam	ACT Math	0.9030738
Integrated Math II EOC Exam	ACT Math	0.9206229

Results, Specific Point Metrics

The metrics for individual scale scores are examined, including both those identified as the optimal cut points that are maximally informed, and the scale score that identifies students as "proficient" on the EOC exam (score of 700).

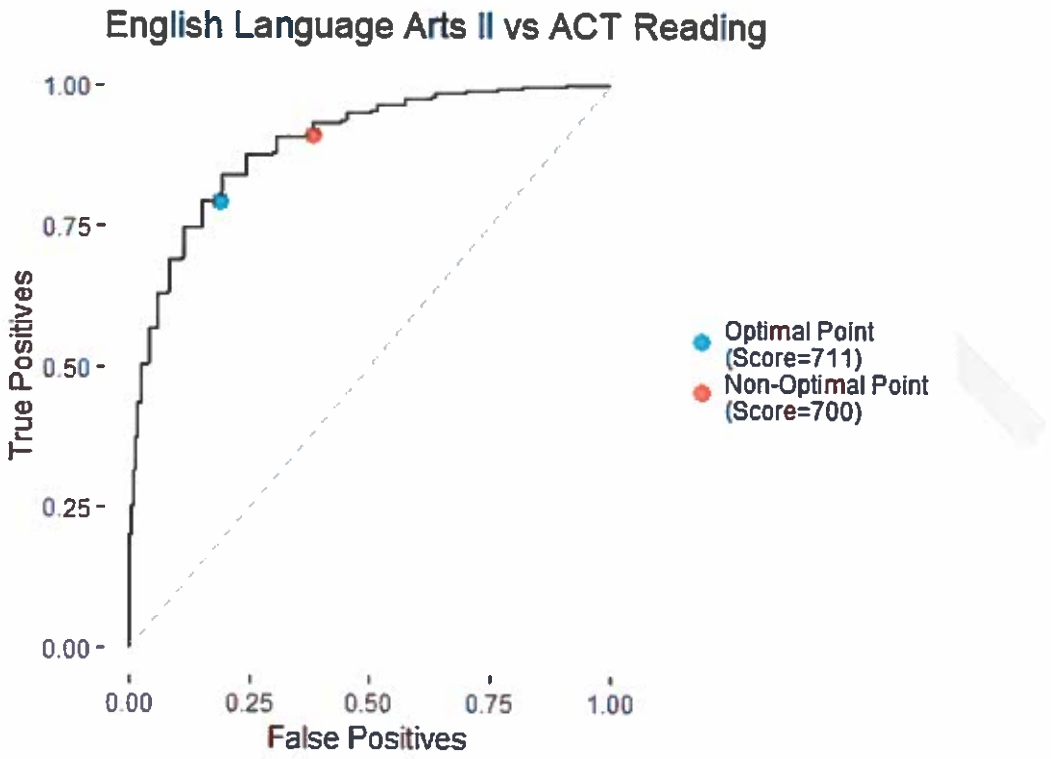
English Language Arts II vs ACT English

Cutoff	Description	True Positive Rate	False Positive Rate	Accuracy	Informedness
708	Optimal Point	0.7755	0.1905	0.7911	0.585
700	Non-Optimal Point	0.8691	0.3206	0.782	0.5484



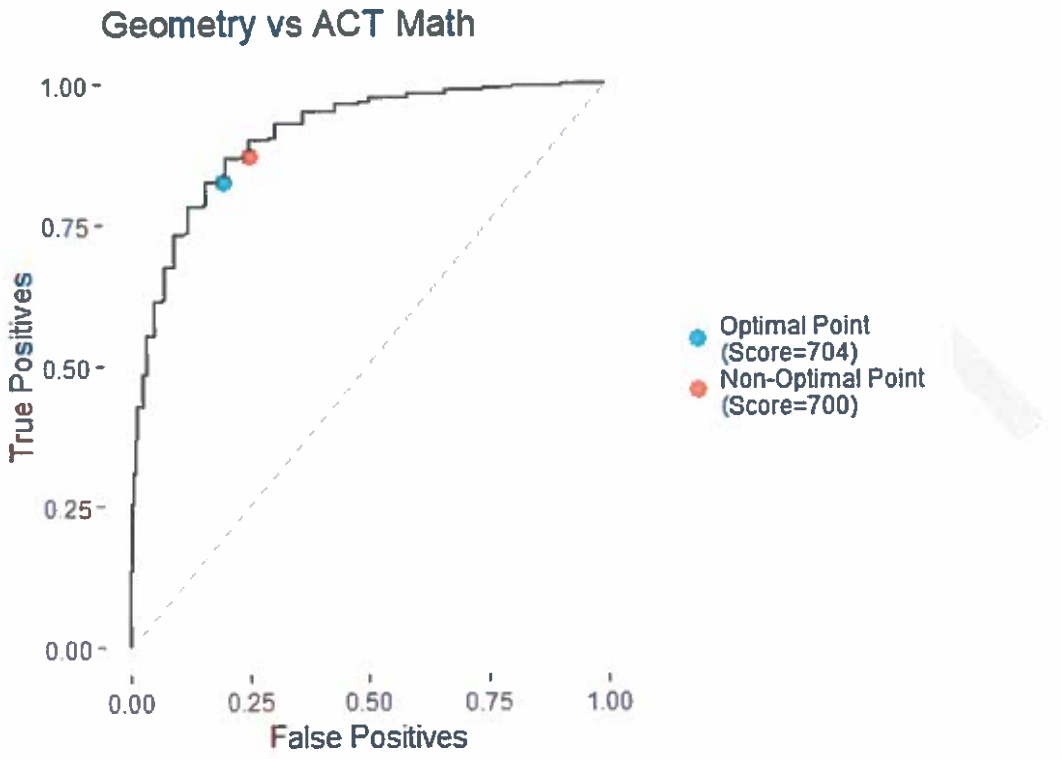
English Language Arts II vs ACT Reading

Cutoff	Description	True Positive Rate	False Positive Rate	Accuracy	Informedness
711	Optimal Point	0.7936	0.1887	0.8033	0.6049
700	Non-Optimal Point	0.9111	0.3792	0.751	0.5319



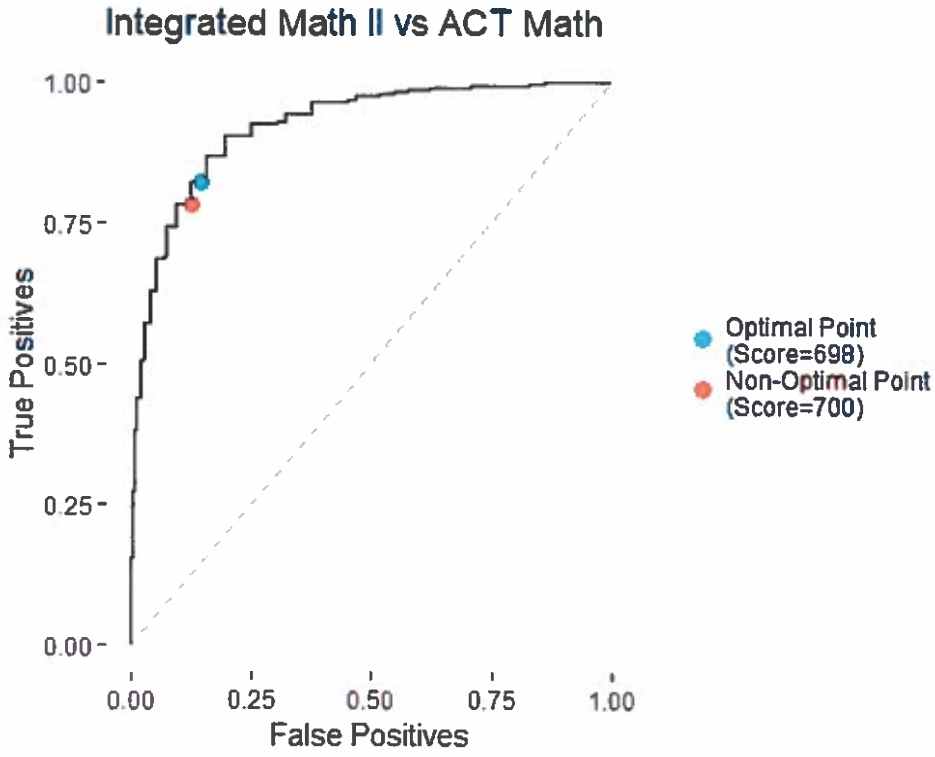
Geometry vs ACT Math

Cutoff	Description	True Positive Rate	False Positive Rate	Accuracy	Informedness
704	Optimal Point	0.825	0.1996	0.806	0.6254
700	Non-Optimal Point	0.8704	0.2538	0.7744	0.6166



Integrated Math II vs ACT Math

Cutoff	Description	True Positive Rate	False Positive Rate	Accuracy	Informedness
698	Optimal Point	0.82	0.1474	0.8465	0.6726
700	Non-Optimal Point	0.7821	0.1256	0.8574	0.6566



Analysis and Conclusion

Both the linear regression R-squared values and the ROC curves's AUC establish a fairly strong relationship between End-Of-Course exams and their respective ACT sub-scores. However, this is for establishing a general relationship between our predictors and outcome variables—if we want to establish optimal cut-scores and examine proposed cut scores, we need to look deeply at the metrics around individual cut scores, whether optimal or not.

In general, we see that the optimal points for the math ROC curves exhibit greater accuracy and more informedness than the English models. For the Integrated Math II model, the optimal point lies *below* 700. If we wanted to consider how well proficient on the Integrated Math II exam is indicative of remediation-free on the ACT Math subsection, then we would need to *raise* the threshold, which lowers both the true positive rate and the false positive rate. Doing so leaves us with a more conservative model that would under-identify students, but would be a trade-off, as we would be more sure that the student does not need remediation.

However, changing the Geometry threshold would be non-ideal in the opposite direction, as this would *increase* the false positive rate, meaning that many more students needing remediation based on their ACT scores would not be identified for remediation based on their EOC exam. This is even more pronounced with the English ROC curves; changing the threshold for the ELA II EOC exam to 700 for both the ACT English and ACT Reading models significantly reduces the accuracy and informedness of the English models and increases the false positive rate.

To conclude, while the optimal points for these models are near 700, forcing the 700 thresholds as a proxy for remediation-free on the ACT likely will introduce more misclassifications. This makes sense, as the two tests are likely not aligned to one another. As such, it is possible a "proficient" score on one (e.g. representing mastery of Ohio's Learning Standards) may not directly equate to "proficient" on another (e.g. representing the ACT's definition of college/career readiness). Looking at the alignment of these tests may be fruitful in understanding both the similarities and differences between these two tests.