ABSTRACT

Rapid changes to statewide achievement tests may pose difficulties for educators tasked with identifying students in need of additional support. The purpose of this study was to evaluate the stability of decision-making accuracy estimates across changes to the statewide achievement test. We analyzed extant data from more than 2,600 students collected in a large suburban district in Wisconsin. Sensitivity and specificity estimates were relatively stable in math. Changes in the criterion measure were associated with decreased sensitivity in reading. When facing situations where the statewide achievement test is changed, educators could continue to use existing screening practices until test vendors or state educational agencies can establish cut-scores associated with the newer test. Lowering cut-scores to establish risk (increasing sensitivity while decreasing specificity) may be prudent in reading.

Accuracy of Universal Screening Cut-Scores Across Changes in State Assessments

David A. Klingbeil, Ethan R. Van Norman, Samuel A. Maurice, Amber L. Schramm, & Chris Birr

INTRODUCTION

- In grades 3 through 8, universal screening is often used to predict year-end proficiency on a statewide achievement test, in order to identify students needing additional support (e.g., Fuchs, Fuchs, & Compton, 2010).
- Several states have made recent changes to their statewide achievement tests, and the Every Student Succeeds Act (2015) provided additional flexibility for measuring student proficiency in math or reading.
- Changes to state achievement tests will change the associated decision-making accuracy of a screening measure (VanDerHeyden, 2011). If rapid changes occur, educators may need to interpret screening performance with little guidance from test vendors or State Educational Agencies.

Research Questions

1. What is the stability of decision-making accuracy outcomes, across a change in the criterion-measure, when using vendor-provided cut-scores that were aligned with the previous criterion test?
2. What is the stability of decision-making accuracy outcomes, across a change in the criterion measure, when using vendor-provided cut scores that were aligned with the previous criterion test?
3. What is the decision-making accuracy when using performance on the previous criterion measure to predict performance on the newly developed criterion measure?

METHODS

Data were collected from students in grades 3 through 8 in all five elementary and both middle schools in a large suburban district in WI. The sample included 2,782 students in 2014-2015 and 2,896 students in 2015-2016.

Measures

- *Measures of Academic Progress* (MAP; Northwest Evaluation Association, 2011*)* – administered in the fall of both years.
- *Smarter Balanced Assessment Consortium* (SBAC) – the state test administered in the spring of 2014-2015
- *Wisconsin Forward Exam* – the state test administered in the spring of 2015-2016.

Student performance on the fall MAP was used to predict student performance on the SBAC or the Wisconsin Forward Exam. We used vendor-provided cut scores and locally derived cut-scores to establish risk on the MAP. We compared the sensitivity (SE), specificity (SP), and AUC, stratified by grade and subject area, between the SBAC and Forward exam using a series of two-proportions tests.

RESULTS

The average base rates of failure on the SBAC ELA (25.8%) and Math (27.2%) were lower than on the Forward Exam (ELA = 38.5%, Math 31.6%).

Question 1: Stability of Vendor Provided Cut-Scores
We examined the stability of decision-making accuracy estimates when using the cut-scores recommended by the MAP for the SBAC (NWEA, 2015).

English Language Arts (ELA). The median SE and SP values, across grades 3 through 8, were similar when using the MAP to predict SBAC performance (.66 and .87) or Forward Exam performance (.64 and .91). The SE and SP values were each significantly different in 2 of 6 grades. Differences in the AUC values were not significant.
Math. Median sensitivity and specificity values were .76 and .89, respectively, in 2014-2015, and .83 and .89, respectively, in 2015-2016, across all grades. The SE values were significantly different in 1 of 6 grades. Differences in SP and AUC values were not significant.

Question 2: Stability of Locally Derived Cut-Scores
We examined the stability of decision-making accuracy estimates when using MAP cut-scores that ensured SE ≥ .90 on the SBAC.

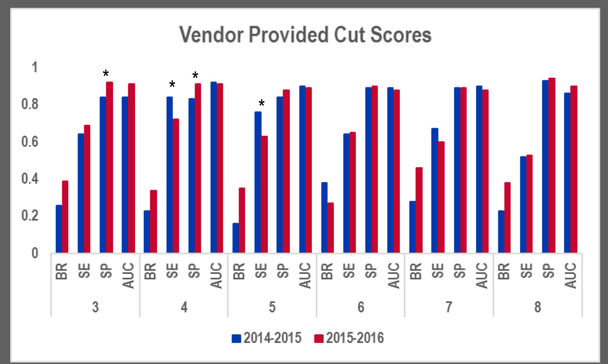
ELA. The median SE and SP values were .92 and .67 when predicting SBAC performance across grades. When using these cut-scores to predict Forward Exam performance, median sensitivity and specificity values were equal to .89 and .71, respectively. The SE and SP values were significantly different in 2 and 3 of 6 grades, respectively. Differences in the AUC values were not significant.
Math. The median SE and SP values were similar when using the MAP to predict SBAC performance (.96 and .71) or Forward Exam performance (.94 and .70). The SE and SP values were significantly different in 1 and 2 of 6 grades, respectively. Differences in the AUC values were not significant.

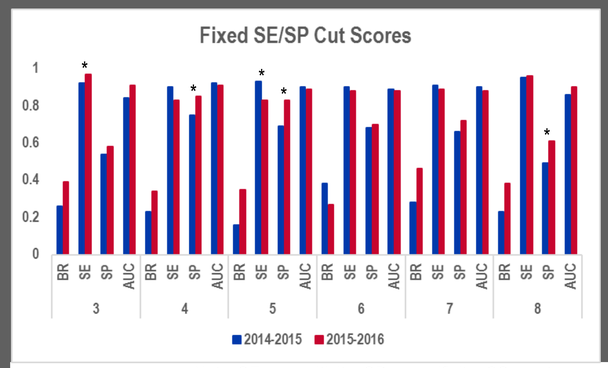
Question 3: Decision-making Accuracy of Previous State-Test
We also investigated the diagnostic accuracy of using SBAC performance to predict Forward Exam proficiency. The cut-score differentiating between with proficiency and basic was used to determine risk.

ELA. Using proficiency status on the SBAC to predict risk on the Forward Exam resulted in a median specificity of .93 but an unacceptable median sensitivity of .63. However, AUC values were generally comparable to those observed with MAP.
Math. Proficiency status on the SBAC resulted in a median specificity of .95. As with ELA, the median sensitivity was unacceptably low .69. AUC values for the SBAC Math were comparable to MAP.

ELA Figures

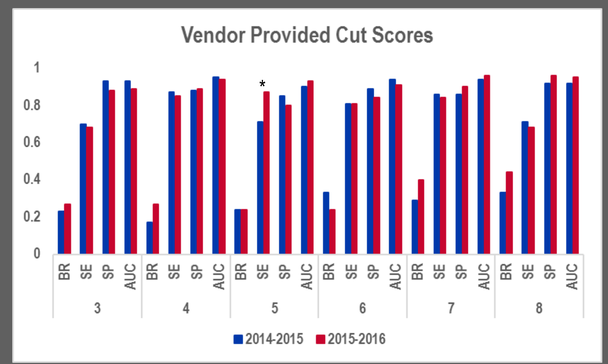
Results of ROC Curve Analysis and Two-Proportions Tests for ELA

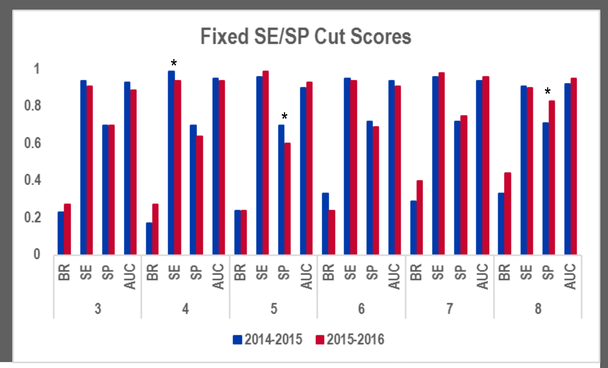




Math Figures

Results of ROC Curve Analysis and Two Proportion Tests for Math





Note. SE = sensitivity; SP= specificity; CS = cut score; BR = base rate; AUC = area under the curve; * = $p < .001$

CONCLUSIONS

- These findings provide initial evidence for educators to use cut-scores that were derived for the previous state test until screening measures can be aligned with the new state test. Sensitivity and specificity estimates were generally stable or improved in math. In reading, sensitivity estimates were either stable or decreased. A conservative approach could be to lower the cut-scores (i.e., provide more students with additional support) to minimize the number of false negatives.
- This study provided further support (e.g. Nelson, Van Norman, & VanDerHeyden, 2016) for the use of local cut-scores over the use of vendor-provided cut scores. The most accurate approach for predicting risk on the Forward Exam was to use locally derived cut-scores that would result in a minimum sensitivity of .90 and specificity of .70 on the SBAC.
- This study provides partial support for the use of prior year state test scores as a screening tool based on AUC values. The threshold for determining risk must be set at a value that will result in acceptable SE values. It is unclear whether classification cut-points (e.g., advanced vs. proficient; basic vs. below basic) will suffice.
- The extent to which these results generalize to other screening measures or state-tests is unknown. Further research is warranted.

Elmbrook Schools — School District of Elmbrook
University of Wisconsin Milwaukee
Georgia State University