Research Brief

Predictive Validity and Forecasting Accuracy for the 2015-16 School Year: AzMERIT

by Christine Burnham, Ph.D.
Assessment Technology Incorporated

Overview: ATI investigates the predictive validity of Galileo® assessments and the forecasting accuracy of Galileo risk levels on an annual basis once districts'/charters' statewide assessment data for individual students has been uploaded into the Galileo database. ATI evaluates predictive validity by examining the correlation between student scores on each district/charter-wide assessment and student scores on the statewide assessment. ATI evaluates forecasting accuracy by examining how students classified at different levels of risk ultimately performed on the statewide assessment. This document provides a comprehensive summary of the research on both the predictive validity of Galileo assessments administered in Arizona during the 2015-16 school year and the forecasting accuracy of Galileo risk levels based on student performance on these assessments with regard to their eventual performance on the AzMERIT assessment. For districts and charters who uploaded their 2016 AzMERIT assessment data, the results of the specific investigations for their administered assessments are available in the online Galileo *Forecast Report*.

Sample: For the purpose of this brief, the first 32 districts/charters to provide ATI with their 2016 AzMERIT data for individual students in grades two through high school in math and reading/English language arts were included in the sample. The 32 districts/charters included in the sample represent students in Arizona and administered 1,594 district/charter-wide assessments in these grades and content areas. The sample included 600,705 scores recorded by 103,746 students from the 32 Arizona school districts and charter schools. On average, each student contributed 5.8 scores to the analyses, which is consistent with an average of 3 district/charter-wide assessments administered per year in each of the two content areas.

Student Performance Measures: The statewide assessment data uploaded by districts/charters contains a scale score for each student as well as an indication of whether the student passed the statewide assessment. For each district/charter-wide assessment administered, ATI performs an Item Response Theory (IRT) analysis which produces a scale score for each student, the Developmental Level (DL) score. Each student is also classified as to their level of risk of failing the statewide assessment based on their performance on all the district/charter-wide assessments they have taken within a given school year. In order of highest to lowest risk of failing the statewide assessment, the possible risk levels comprise "High Risk," "Moderate Risk," "Low Risk," and "On Course."

Galileo Assessments and Timeframes: The 1,594 district/charter-wide assessments were not identical across all district/charters, nor were there constraints regarding the time of year during which the assessments were administered to students. Many of the assessments were standard Galileo Comprehensive Benchmark Assessment Series (CBAS) assessments, whereas others were district/charter curriculum-aligned assessments. Student DL scores on Galileo assessments within the same grade level and subject are on a common scale established through IRT analyses and are comparable, even if there are differences in the content of the assessments. Therefore, the predictive validity and forecast accuracy analyses were conducted on the pool of student Galileo scores for each grade level and subject regardless of the specific assessment that was administered. However, it would be inappropriate to include student scores from the beginning, middle, and end of the year in a single correlation analysis, because it is expected that student DL scores will increase as the year progresses, and this increase would obscure any correlation between student DL scores and end-of-year AzMERIT scores. Therefore, six separate predictive validity correlation analyses were conducted for each grade level and subject: one analysis for each of 6 timeframes. The date ranges for each student score timeframe is listed in Table 1 below

TABLE 1
Date ranges for the six timeframes used in the correlation study

Timeframe	Start Date	End Date
T1	7/1/2015	9/15/2015
T2	9/16/2015	11/30/2015
T3	12/1/2015	1/15/2016
T4	1/16/2016	2/15/2016
T5	2/16/2016	4/15/2016
T6	4/16/2016	6/30/2016

Predictive Validity Analyses: Predictive validity analyses examine the strength of the relationship between two measures of student performance, in this case the student DL scores on an assessment in a given grade and content area and the student scores on the statewide assessment in the same grade and content area. Predictive validity analyses can produce correlation statistics that range from -1 to +1, although typically only positive values are observed in this context. A positive correlation indicates a positive relationship, that is high scores on one measure are associated with high scores on the other measure. A negative correlation would indicate a negative relationship, that is high scores on one measure are associated with low scores on the other measure. A correlation of zero would indicate no relationship. Values of positive or negative one indicate a perfect relationship between the two measures and are rarely observed under real-world circumstances. The predictive validity analysis for each grade level, subject, and timeframe were performed on the pooled set of student scores for all of the district/charter-wide assessments administered by the group of 32 districts/charters in the relevant grades and subjects during the 2016-17 school year.

Predictive Validity Results: Table 2 illustrates the correlation observed for the student Galileo® DL scores on the assessments administered in each grade and content area during each timeframe. As the chart shows, the correlations range, on average, from 0.70 for timeframe 1 to 0.78 for timeframe 6 with an overall mean of 0.74. The data demonstrates a tendency for correlation values to increase as the year progresses, reflecting the fact that assessments taken early in the year can capture student ability levels before they have begun to acquire knowledge and skills in the content area. This is especially true of high school math. A correlation between 0.7 and 0.9 indicates a high correlation between the two measures, while a correlation between 0.5 and 0.7 indicates a moderate correlation. Thus, the observed correlations suggest that student scores on the 2015-16 Galileo assessments were generally strongly related to student scores on the 2016 statewide assessment.

TABLE 2
Mean correlations between scores for the 2015-16 Galileo assessments and scores for the 2016 statewide assessment for each grade level, content area, and timeframe.

		Timeframe						
Subject	**Grade**	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**Average**
ELA	3	0.78	0.77	0.79	0.73	0.80	0.81	**0.78**
ELA	4	0.76	0.75	0.79	0.75	0.80	0.78	**0.78**
ELA	5	0.80	0.73	0.80	0.75	0.74	0.81	**0.77**
ELA	6	0.80	0.78	0.78	0.75	0.79	0.80	**0.78**
ELA	7	0.79	0.73	0.80	0.78	0.74	0.81	**0.77**
ELA	8	0.75	0.74	0.77	0.79	0.76	0.81	**0.77**
ELA	9	0.59	0.76	0.77	0.71	0.69	0.75	**0.71**
ELA	10	0.67	0.65	0.69	0.76	0.68	0.70	**0.69**
ELA	11	0.70	0.66	0.59	0.77	0.66	0.67	**0.68**
Math	3	0.68	0.70	0.77	0.72	0.75	0.84	**0.74**
Math	4	0.76	0.70	0.77	0.79	0.75	0.83	**0.77**
Math	5	0.74	0.75	0.79	0.79	0.77	0.85	**0.78**
Math	6	0.74	0.75	0.80	0.80	0.82	0.83	**0.79**
Math	7	0.74	0.74	0.79	0.83	0.79	0.85	**0.79**
Math	8	0.64	0.56	0.74	0.77	0.71	0.82	**0.71**
Algebra I	HS	0.52	0.58	0.68	0.71	0.64	0.73	**0.64**
Geometry	HS	0.52	0.65	0.58	0.74	0.69	0.71	**0.65**
Algebra II	HS	0.65	0.72	0.72	0.63	0.63	0.71	**0.68**
Average of Correlations		**0.70**	**0.71**	**0.75**	**0.75**	**0.73**	**0.78**	**0.74**

Forecasting Accuracy Analyses: Forecasting accuracy analyses examine the accuracy with which Galileo® risk levels for individual students predicted their ultimate performance on the relevant statewide assessment. Risk levels provide an indication of the likelihood that a student is at risk to fail the statewide assessment. Although risk levels represent a continuum of risk, for the purpose of forecasting accuracy analyses, students who are classified as "On Course" or as "Low Risk" are predicted to pass the statewide assessment while students who are classified as "Moderate Risk" or "High Risk" are predicted to fail the statewide assessment. Forecasting accuracy analyses were conducted for the group of 32 districts/charters described previously. The pooled student score data set from these 32 districts/charters included risk level classifications and the corresponding AzMERIT classification for 173,795 student/content area records (most students contributed two records: one for math and one for ELA).

Forecasting Accuracy Results: Figure 1 illustrates the percentage of students in each risk level who passed the statewide assessment. Figure 2 illustrates the overall forecasting accuracy as well as the forecasting accuracy for each risk level. There are three important aspects of the forecasting accuracy analysis to evaluate. First, as student risk level decreases the likelihood of success on the statewide assessment should increase. This is a prerequisite for accurate forecasting. As Figure 1 shows, the majority of students who were classified as being "On Course" based on their performance on the

Galileo® district/charter-wide assessments did in fact pass the statewide assessment, while the majority of those who had been classified as being at "High Risk" of not demonstrating mastery on the statewide assessment did in fact fail. The other two risk level groups performed as expected as well.

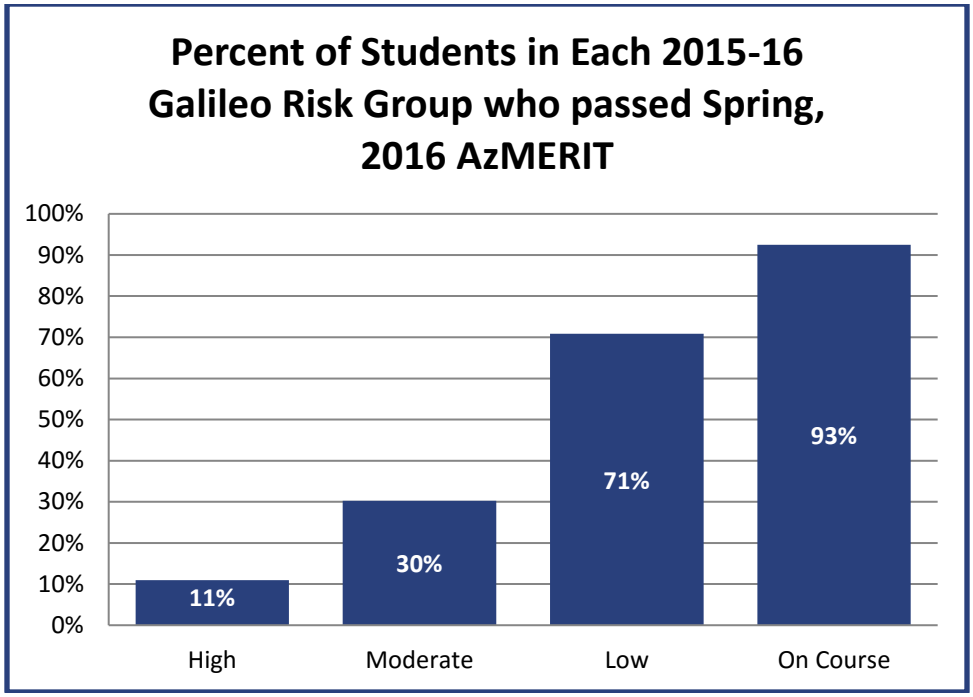

Figure 1
Mean percentage of students passing the statewide assessment for each risk level.

Second, overall forecasting accuracy should be adequately high. ATI considers forecasting accuracy to be adequate if a student's risk level accurately predicted performance on the statewide assessment for at least 75 percent of students within a district/charter. As Figure 2 shows, the overall forecasting accuracy was quite high, with statewide test performance accurately forecast, on average, for 82 percent of students.

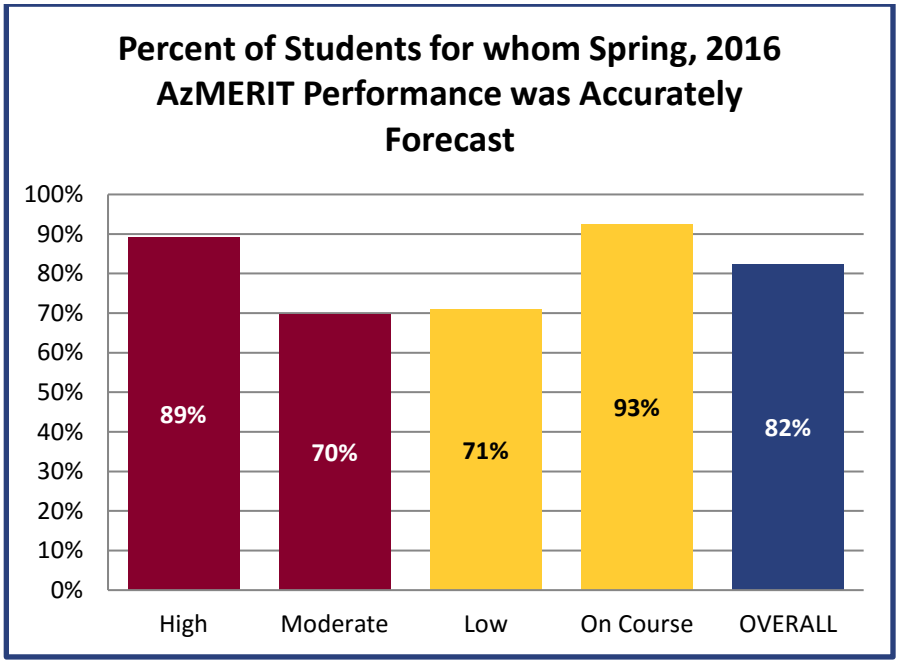

Figure 2
Overall forecasting accuracy and forecasting accuracy for each risk level.

Third, forecasting accuracy should be highest in cases where student performance is most consistent. Students who consistently perform well on Galileo® assessments and are thus classified as "On Course" should consistently pass the statewide assessment. Conversely, students who consistently perform poorly on Galileo assessments and are classified as "High Risk" should consistently fail to pass the statewide assessment. Students whose performance on Galileo assessments is more variable (i.e., the "bubble" students who sometimes perform well and sometimes don't) should also display more variable performance on the statewide assessment. As Figure 2 shows, and as expected, forecasting accuracy was highest for students classified as "On Course" and "High Risk" and somewhat lower for students classified as "Low Risk" and "Moderate Risk." It should be noted that, if teachers and administrators are using the data provided by Galileo district/charter-wide assessments to implement effective interventions, many students who have been classified as being at some risk of failing the statewide assessment should pass it instead, thereby reducing the accuracy of risk assessment forecasts for the those student groups. ATI therefore considers a certain degree of inaccuracy in predictions of failure to be a sign of success.

Conclusion: The research presented in this document was conducted to evaluate predictive validity and forecasting accuracy for the 2015-16 school year. The results suggest that the 2015-16 Galileo assessments demonstrated adequate levels of predictive validity. The results also suggest that the 2015-16 Galileo risk levels displayed adequate levels of accuracy in forecasting student performance on the statewide assessment. This research is consistent with similar research investigations performed in previous years and suggests that Galileo assessments and risk levels continue to demonstrate adequate levels of predictive validity and forecasting accuracy.