Galileo® K-12 Online Psychometrics FAQ

How does ATI calculate my district's/charter school's psychometrics benchmark (i.e., district-wide) test data?

To generate ATI's scaled scores (ATI refers to these as the Developmental Level Scores or DL scores) ATI uses an analysis based in Item Response Theory (IRT). IRT takes information about the difficulty of the items into account when generating the estimates of student ability (i.e., the DL scores). What this means for the district/charter school is that a change in DL scores is a direct measure of growth. For example, if a student obtains a DL score of 1000 on district-wide test #1 and 1100 on district-wide test #2, that would mean that the student's ability increased by 100 points or one standard deviation. In contrast, by looking at raw scores or percent correct, you cannot be sure what it means if scores increased. For example, if a student obtained 70% correct on district-wide test #1 and 80% on district-wide test #2, this increase might be related to the fact that the items on district-wide test #2 were easier or the fact that the student's ability increased, or both.

So DL scores are generally preferable to raw scores as a way to evaluate student progress; however, sometimes the results of this approach are not so intuitive. For example, a teacher may see that his or her class improved in terms of percent correct from district-wide test #1 to district-wide test #2, but that their DL scores have actually decreased. In this case, the items on district-wide test #2 were probably easier than those on district-wide test #1, so essentially the students didn't get as much "credit" for getting them right.

Once ATI has the DL scores, ATI sets the cut scores that correspond to the performance levels (e.g., Highly Proficient, Proficient, Partially Proficient, and Minimally Proficient). ATI uses two approaches to set cut scores. The first approach is called ***equipercentile equating****.*

Under the equipercentile equating approach, ATI aligns the distribution of student scores on the district-wide assessment to the distribution of scores on the analogous state assessment for the same district/charter school (i.e., we align the 3rd grade math district-wide test to the most recent 3rd grade state standardized assessment, such as AzMERIT in Arizona). We identify the percentile ranks at which students in the district/charter school attained the various cut scores on the state standardized assessment, identify the same percentile ranks in the distribution of scores on the district-wide assessment, and the DL scores at those percentile ranks become the cut scores on the district-wide assessment. This allows ATI to identify not only how many students are likely to pass the state standardized assessment at the end of the year, but ***which*** *students. Analyses of the accuracy of forecasting indicate that the equipercentile equating approach is highly accurate in forecasting which students are likely to show mastery on the state test and which are not.*

Although the equipercentile equating approach is very accurate, it makes it difficult to detect changes in district-wide performance throughout the year. Individual students, teachers, and schools can improve their performance, but it is not possible for the entire group to show much improvement across district-wide tests since the percentages in each performance level on each district-wide test are always fixed to the state standardized assessment percentages from the previous year.

For that reason, ATI uses another approach, the ***growth approach****, which is applied to district-wide tests #2, #3, etc. With the growth approach, ATI increases the cut scores from one district-wide test to another by the amount of expected growth for that time period for that grade and content area. The expected growth is calculated based on the average amount of growth seen in the past year for each grade and content area across all the districts/charter schools utilizing Galileo.*

So, for example, a 4th grade Math #1 was administered in October and a 4th grade Math #2 was administered in December. If the cut score for Proficient on 4th grade Math #1 was 722 and the expected growth between October and December was 19 points, then ATI raises the cut score for Proficient on 4th grade Math #2 to 741. This allows teachers to see whether students are maintaining the status quo by showing only the expected growth, showing more growth than expected, or showing less growth than expected. We have also found that the growth approach is very accurate in forecasting student performance on the state standardized assessment.

The growth approach relies on the accuracy of the alignment between the distribution of student scores on the initial test used to establish the baseline and the distribution of student scores on state standardized assessments. Therefore, it is important that this baseline is reliable and as accurate a reflection of initial student ability as possible.

ATI always uses equipercentile equating approach for the first test administered during the year. In some cases where a pre-test is administered, ATI may also use equipercentile equating for district-wide test #1. In grades and content areas where data related to expected growth is available (e.g., high school algebra), ATI begins using the growth approach for subsequent tests. In grades and content areas where expected growth data is not available, ATI uses the equipercentile equating approach for each test.

Why is it important to have the majority of the students take the assessment for psychometrics purposes?

ATI tries to match up the distribution of student scores on the district-wide test to the district's/charter school's state standardized assessment's distribution. Therefore, ATI wants the best sample of the student population who will be taking the state standardized assessment. Generally, we like to see that more than 85% of students who will take the state standardized assessment have been tested.

If a district/charter school decided to administer a district-wide test to a subset of the student population such as only all special education students and ATI

aligned the test data to their entire district/charter school state standardized assessment distribution, this would be inaccurate because it would not be the same population of students.

If a district/charter school decided to administer a district-wide test to just one of their schools, ATI would need to know this information. If possible, we will use the state standardized assessment distribution for that one school for alignment. You should discuss your plans with your ATI Field Services Coordinator (800.367.4762) before giving a test to anything less than the full population of students who will take the statewide test in that grade and content area.

Why does ATI need to have my district's/charter school's state standardized assessment data?

ATI utilizes your district's/charter school's state standardized assessment data to "align" this data to your district-wide test data and generate cut scores for the assessments. Generally, ATI can get this data from the state, but in some cases you may be asked to provide it. If your district/charter school is unable to provide this information and you administer the Comprehensive Blueprint Assessment Series (CBAS) assessments, ATI can also use the average cut scores from all districts/charter schools who administered the CBAS for the grade and content area.

How can you compare the results from one district-wide test to another when the questions and/or standards are not the same?

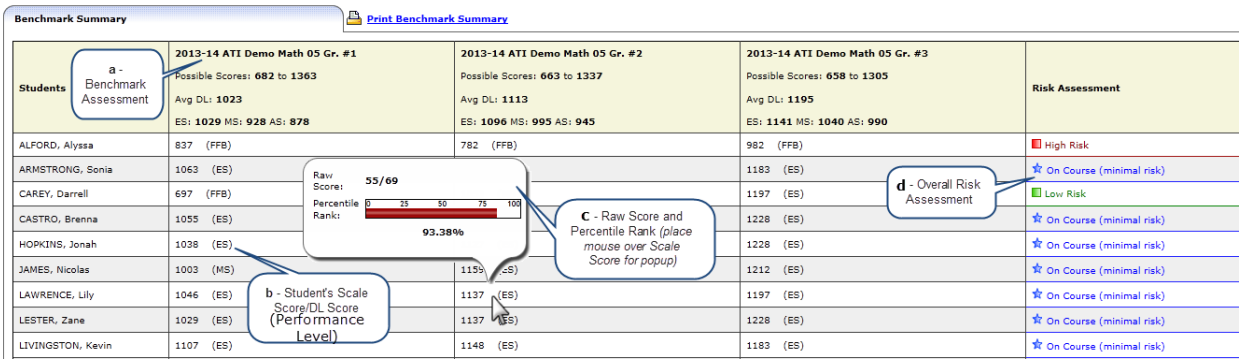
Over time ATI has collected data on how students perform on the test items (e.g., discrimination, difficulty, and guessing parameter values). With this information, ATI can accurately estimate a student's ability based on how they respond to any set of items. The difficulty estimates for the items in the Galileo item banks are all on the same scale within each grade level and content area. Therefore, ATI is able to compare the results from one district-wide test to another district-wide test even if they involve different sets of items, because any subset of items will contain difficulty estimates that are on the common item difficulty scale for the grade and content area. This means that the DL scores that are informed by the item difficulty levels are placed on a common scale and can be compared across tests. A District-level user can generate the Item Parameters report to view the item parameter estimates for each item on a district-wide test.

What is a DL score and how do I use it?

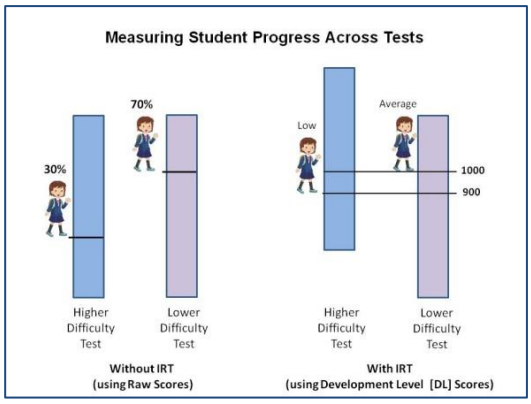
DL is the abbreviation for Developmental Level. The DL score is a scale score similar to the scale score a student receives on the statewide test. A teacher should use the student's DL scores instead of the student's raw scores as a way to evaluate student progress because DL scores take into account the difficulty of the items on each test.

What is the difference between the raw scores and DL scores? Why are DL scores "better"?

A student's raw score just lets the teacher know that the student had "x" number of correct answers out of the total number of test questions. No factors, such as level of difficulty of the test, are taken into account. The DL score, on the other hand, takes into account the total number of questions the student answered correctly and the level of difficulty. For this reason, a teacher should use the student's DL score instead of the student's raw scores as a way to evaluate student progress.



A student took two tests and obtained a raw score of 30% on the first and 70% on the second. A change in raw scores would be a true indication of growth <u>IF and ONLY IF</u> the items on each test were identical in terms of difficulty. But without an IRT analysis we don't know if the items on the tests were of equal difficulty. Therefore, we cannot interpret this increase in raw scores as an indication of growth.



With an IRT analysis, we have information about the difficulty of each item on each test. We take this into account when we generate the estimates of student ability (DL Scores), making the DL scores a true indication of growth.

For example, if a student obtained a DL score of 1000 on District-wide test #1 and 1100 on District-wide test #2, that would mean that the student's ability increased by 100 points or one standard deviation. In contrast, by looking at raw scores or percent correct, we can't be sure what it means if scores increased. For example, if a student obtains 30% correct on District-wide test #1 and 70% on District-wide test #2, this increase might be related to the fact that the items on District-wide test #2 were easier or the fact that the student's ability increased, or both.

How are the various risk levels (high risk, moderate risk, low risk, and on course) assigned?

The risk level is based on student performance across all district-wide assessments administered. The risk level is refined as additional tests are administered and more information about student performance is gained.

	Administered District-wide Test		
Possible Risk Levels	***First Test***	***Second Test***	***Third Test***
High Risk			✓
Moderate Risk	✓	✓	✓
Low Risk		✓	✓
On Course	✓	✓	✓

The risk levels are based on whether the student scored above or below the cut score for proficient (or "Proficient", in AZ) on each district-wide assessment that was taken. This table shows the risk level classifications based on the number of district-wide assessments on which the student scored below the cut score for Proficient.

	# of District-wide Assessments Taken		
# of district-wide scores below the cut score for Proficient	***One Test***	***Two Tests***	***Three Tests***
3			*High*
2		*Moderate*	*Moderate*
1	*Moderate*	*Low*	*Low*
0	*On Course*	*On Course*	*On Course*

What is the best prediction of student performance on the statewide test?

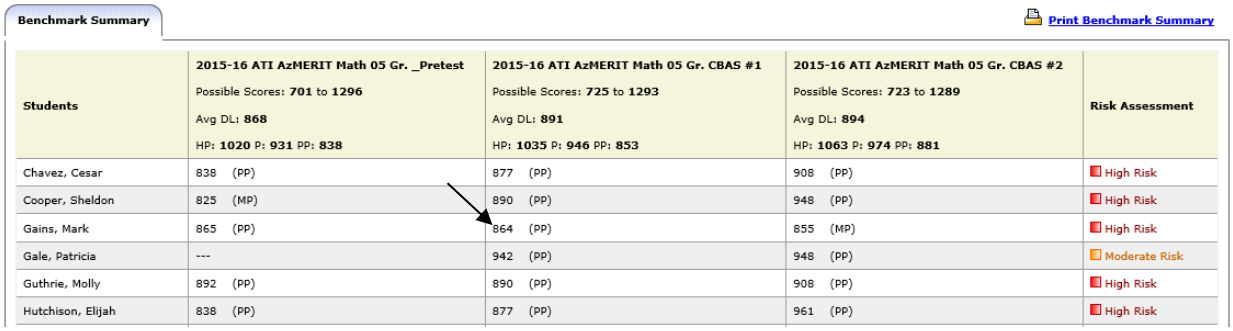
You should use the risk level because it takes into account information from multiple tests.

What would be considered a good (passing) score?

Each test has a cut score that represents the DL score students need to have obtained in order to not be at risk for not passing the state standardized assessment. Although this cut score will generally correspond to a raw score, you should use the student's DL score/performance level and not their raw score to determine if a student is passing. The student's performance level and their DL score takes into account the level of difficulty for the test; a raw score does not.

In this example for CBAS #1 test, Mark Gains has a DL score of 864. The cut score of 946 is the "Proficient" in Arizona. Any DL score of 946 or above would

indicate that the student is on target for passing the state standardized assessment.



Benchmark Summary

Print Benchmark Summary

Students	2015-16 ATI AzMERIT Math 05 Gr. _Pretest Possible Scores: 701 to 1296 Avg DL: 868 HP: 1020 P: 931 PP: 838	2015-16 ATI AzMERIT Math 05 Gr. CBAS #1 Possible Scores: 725 to 1293 Avg DL: 891 HP: 1035 P: 946 PP: 853	2015-16 ATI AzMERIT Math 05 Gr. CBAS #2 Possible Scores: 723 to 1289 Avg DL: 894 HP: 1063 P: 974 PP: 881	Risk Assessment
Chavez, Cesar	838 (PP)	877 (PP)	908 (PP)	High Risk
Cooper, Sheldon	825 (MP)	890 (PP)	948 (PP)	High Risk
Gains, Mark	865 (PP)	864 (PP)	855 (MP)	High Risk
Gale, Patricia	---	942 (PP)	948 (PP)	Moderate Risk
Guthrie, Molly	892 (PP)	890 (PP)	908 (PP)	High Risk
Hutchison, Elijah	838 (PP)	877 (PP)	961 (PP)	High Risk

What does it mean if a student is classified as passing a test (e.g., Proficient")? Does it mean that student would pass the state standardized assessment if they took it today?

No. It means they are likely to pass the state standardized assessment at the end of the year provided they continue on the same path they are currently on.

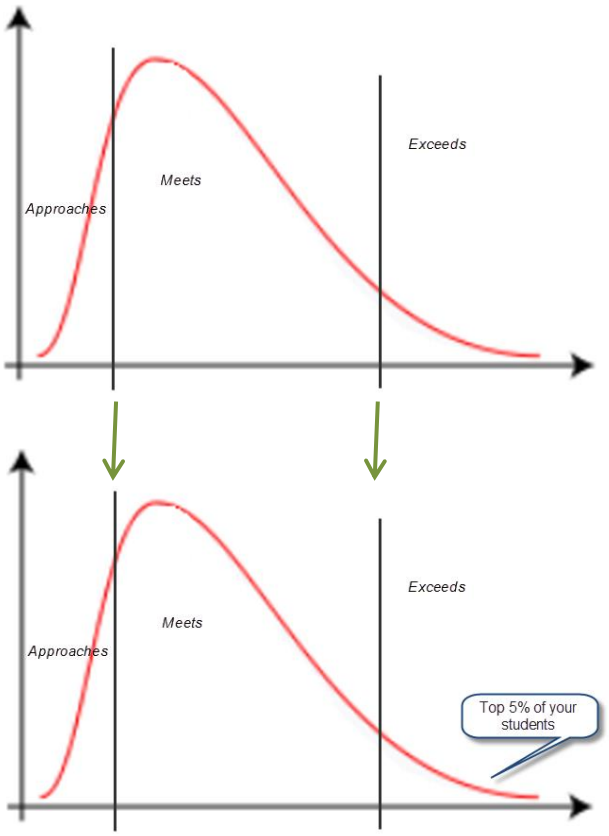
How can a student be designated as passing in terms of performance level if they have a low raw score? Or as not passing if they have a high raw score?

The student's raw score reflects only the number of questions they answered correctly out of the total number of test questions. The DL score reflects the total number of items answered correctly, the difficulty of the test, and the student's performance relative to other students in the district/charter school. Instead of looking at the student's raw score, the teacher should look at the student's DL score.

Why did a student need to get "x" raw score to be classified as a certain performance level (e.g., Highly Proficient, Proficient)?

ATI takes all the district/charter school students who took the district-wide test and arranges them into a distribution of current scores. ATI reviews the district's/charter school's state standardized assessment data and identifies the percentage of students who fall into each of the state performance level. ATI then finds the district-wide score that closely aligns to each performance level.

For example, all the students in a district/charter school take the district-wide test. ATI calculates the students' DL scores. If this is a pretest or a district-wide test #1, ATI looks at the district's/charter school's state standardized assessment data and identifies the percent of students that were classified as Exceeds last year in that grade and content area (e.g., 5%). ATI then locates the raw score/DL score that will identify the top 5% of students on the district-wide test. That raw score/DL score becomes the cut score for Exceeds on the district-wide test. This is repeated for each of the state standardized assessment performance levels.

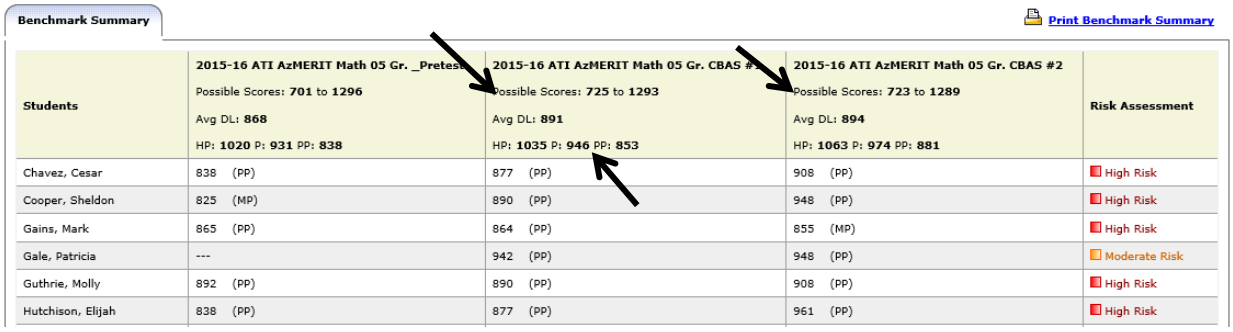


Your state standardized test data

ATI locates the cut score for each performance level based on your state standardized test data

Do the cut scores change from one test to another?

Yes, the cut scores can change one district-wide test to another district-wide test. This is because you are administering a different test and because students are expected to grow throughout the year.



Benchmark Summary — Print Benchmark Summary

Students	2015-16 ATI AzMERIT Math 05 Gr. _Pretest Possible Scores: 701 to 1296 Avg DL: 868 HP: 1020 P: 931 PP: 838	2015-16 ATI AzMERIT Math 05 Gr. CBAS #1 Possible Scores: 725 to 1293 Avg DL: 891 HP: 1035 P: 946 PP: 853	2015-16 ATI AzMERIT Math 05 Gr. CBAS #2 Possible Scores: 723 to 1289 Avg DL: 894 HP: 1063 P: 974 PP: 881	Risk Assessment
Chavez, Cesar	838 (PP)	877 (PP)	908 (PP)	High Risk
Cooper, Sheldon	825 (MP)	890 (PP)	948 (PP)	High Risk
Gains, Mark	865 (PP)	864 (PP)	855 (MP)	High Risk
Gale, Patricia	---	942 (PP)	948 (PP)	Moderate Risk
Guthrie, Molly	892 (PP)	890 (PP)	908 (PP)	High Risk
Hutchison, Elijah	838 (PP)	877 (PP)	961 (PP)	High Risk

Why are the *Aggregate Multi-Test* and/or *Benchmark Results* reports not available when the test is completed?

In order for ATI to run the IRT analysis and set cut scores, ATI looks for at least an 85% participation rate. Once ATI has the necessary participation rate, the analysis will be run. The Testing Activity report allows administrators to check the progress of testing across each individual school. You should contact your Assessment Technology's Field Services Coordinator or the ATI's Research team

(800.367.4762 or Research@ati-online.com) when testing is at least 85% complete in all schools so that ATI can start the analysis. Districts/charter schools should allow at least ten (10) business days for the analysis.

Your district/charter school Galileo Administrator should also contact Assessment Technology if they do not foresee meeting 85% participation. ATI can then run the analysis on your district's/charter school's data using the current participation rate; however in some cases, the results will need to be interpreted with caution if the full distribution of students was not tested.

It is not necessary for Arizona districts/charter schools to contact ATI once their core subject district-wide testing is completed. This is because ATI automatically generates the analysis on the core subjects' district-wide tests as they are administered. If your district/charter school is within Arizona and testing in non-core subject areas or if you are outside of Arizona, your Galileo Administrator needs to contact ATI's Research department (Research@ati-online.com) to let them know that testing is complete and that IRT analysis can begin.

How accurate are the report forecasts?

The Forecast Report within Galileo provides district/charter school users in AZ, CO, and MA with analysis of the accuracy of ATI's predictions regarding whether students would successfully pass their state standardized assessment. Users in other states are provided with an offline version of this report. This report includes three columns in particular – Percent Accurately Forecast, Percent Accuracy, and Correlations with State Test.

Percent Accurately Forecast: The percent of students whose performance on the state standardized test was consistent with what ATI predicted based on their performance on the district-wide assessments. For the purposes of this report, ATI predicts that students who are classified as "On Course" or "Low Risk" will most likely pass the state standardized assessment, and that students who are classified as "Moderate Risk" or "High Risk" will most likely NOT pass the state standardized assessment. When using the Risk Groups to plan re-teaching and intervention efforts, it should be recognized that "Low Risk" students are at some degree of risk of not passing the state standardized assessment.

Overall Percent Accuracy: ATI's accuracy of forecast of state standardized assessment performance for all tested students in all Risk Groups.

Correlations with State Test: The Pearson correlation coefficient indicating the degree to which student scores on the district-wide assessment are consistent with their scores on the state standardized assessment. In general practice a correlation of 0.5 is considered to be "strong." A correlation of 0.65 or 0.70 is even better.

There is very little growth between two district-wide tests. Why?

When evaluating the amount of growth observed between tests, it is important to consider how much growth was expected. ATI's growth expectations vary

depending on how many days transpired between the dates that the two assessments were administered. The Categorical Growth Analysis provided as part of the Student Growth and Achievement Report compares the observed growth for a group of students to the growth expectation for the time period between the tests. This analysis tells you whether students, as a group, showed significantly more growth than expected (Expected Growth Exceeded), about as much growth as expected (Expected Growth Maintained), or significantly less growth than expected (Expected Growth Not Maintained).

Please note: When tests are taken offline, the test date used to define growth expectations is the date that the answer sheet is scanned into Galileo using Scanline. It is important, therefore, that tests be scanned in the order they are administered: the pretest, test #1, test #2, then the posttest. Scanning should occur as close as possible to the date of the test administration. This ensures that the growth expectations are accurately calculated.

I still have some questions. Who can I contact?

You can contact your ATI Field Services Coordinator at 800.367.4762. You may also view the "The Psyche of Psychometrics" video located on ATI's YouTube channel, www.ati-online.com.