The 2018 Science MCA-III Benchmark Report

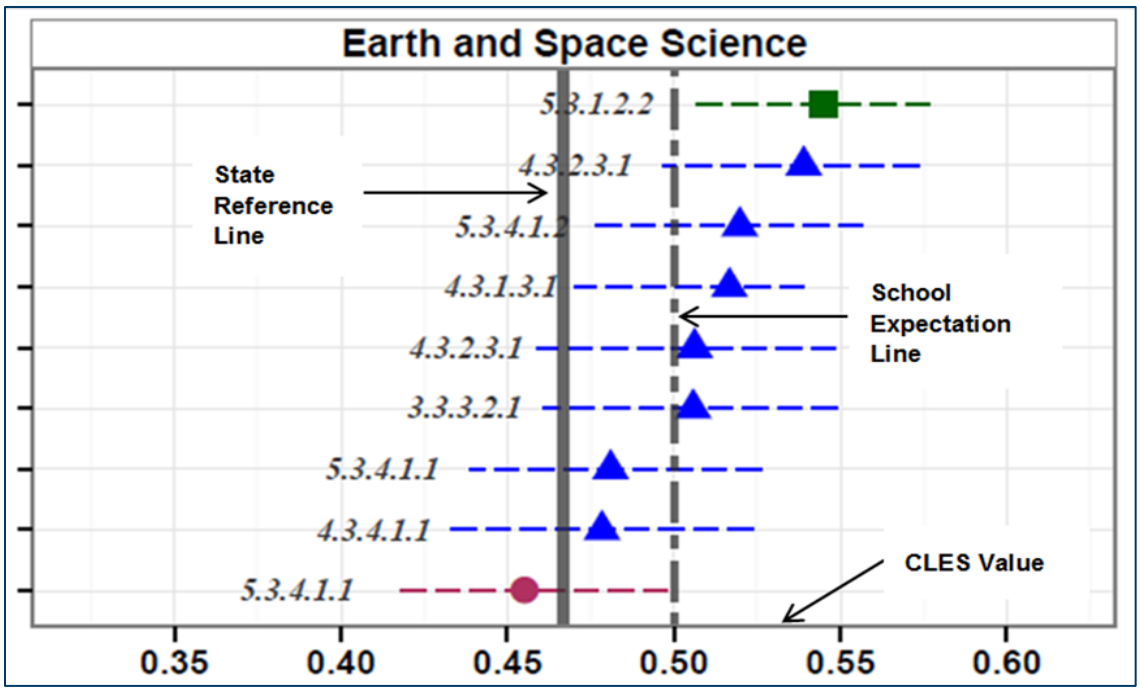
The Science MCA-III Benchmark Report is a tool that educators can use to compare the performance of students in their school on individual MCA items (and their aligned benchmarks) **relative to their overall performance** on the Science MCA-III**.** That is, a school's performance on each MCA item is described in terms of a deviation around the performance expected based on its students' scores on the entire test. The Science reports for 2015 to 2018 use this approach. The Science reports from 2012 to 2014 used a different approach comparing school and state performance on benchmark items. The first section of this document presents an introduction to the benchmark reports and their interpretation. The second section presents a more-detailed discussion of the technical details involved in the calculations that are displayed in the reports.

The Science MCA-III Benchmark Reports are organized by the content strands in the Minnesota Academic Standards for Science (2009). A separate graph is produced for each strand to report school performance on the items from that strand. A list of the benchmarks assessed by the Science MCA-III is included in the document MCA-III Test Specifications: Science available on the Test Specifications page of the MDE website. View the Test Specifications page (MDE > Districts, Schools and Educators > Statewide Testing > Test Specifications).

How to Read the Science MCA-III Benchmark Reports

Figure 1 displays the performance for a school on nine items from the Grades 3-5 Earth and Space Science Strand. Each plotted point represents performance on an item relative to overall performance expectation for the school (represented by the dashed vertical line that crosses the x-axis at value 0.50).

Figure 1. Sample Benchmark Report for the Elementary Earth and Space Science Strand.



The relative performance on individual items for students within a school is reported using the Common Language Effect Size (CLES). When item performance for a group of students is equivalent to that expected given their overall MCA scores, the CLES will equal 0.50. CLES values greater than 0.50 indicate performance on the item by students at the school exceeds that expected from their overall scores. CLES values less than 0.50 have the opposite implication: school performance on the item is lower than expected based on the overall MCA scores of students in the school.

In addition to markers showing the school's relative performance on each item, a synthetic state-wide reference line (the solid gray vertical line) is included in the report to add a normative perspective. This state-wide reference line reflects the performance expected from students at the state mean in overall science ability compared to that expected from students in the school across *all* items. Thus, the CLES value associated with state vs. school reference lines reflects global, rather than benchmark or strand differences in performance expectations. When the CLES value for the state reference line is greater than 0.50, it indicates overall expected state performance is greater than expected school performance; values less than 0.50 have the opposite meaning.

Benchmark Indicators

Individual item benchmarks within each strand are identified by a 5-digit code. Relative performance on each item is indicated by a color and shape coded symbol and a dashed line extending from the symbol. The symbol's horizontal position indicates the actual effect size value for the item (on the CLES metric). The dashed line around the symbol represents a corresponding 95% credible interval (i.e., a 0.95 probability range of plausible CLES values given the data). Within each strand, the items are arranged from highest relative performance at the top right of the graphic to lowest relative performance at the bottom left. As described in Table 1, the color and shape of each plotted symbol indicate how the school's students performed on the item **relative to expectations** based on their overall Science MCA-III scores. Because the state reference line is based on comparison of state vs. school expectation across *all* items, comparison of individual benchmark performance with the state reference line is not appropriate.

Table 1. Item Marker Color and Shape Codes

Green Square: Students performed significantly above expectation on the item. Assigned to markers to the *right* of the dashed vertical line with credible bands that do not overlap the line.
Blue Triangle: Students performed near expectation on the item. Assigned to markers with credible bands that overlap the dashed vertical line.
Red Circle: Students performed significantly below expectation on the item. Assigned to markers to the *left* of the dashed vertical line with credible bands that do not overlap the line.

Evaluating Performance Differences between Benchmarks

In making comparisons between pairs of items within a strand, pay close attention to the amount of overlap of the credible bands for those items. If their credible bands overlap by more than one-half, regardless of color or position of the markers, performance on those items may be considered statistically equivalent. In other words, if the bands on two different items have substantial overlap, there is little credible evidence to suggest that actual performance was significantly different on the two items. If the credible bands across two items do not overlap, then there is very clear evidence of a reliable difference in performance between the two items.

Cautions in Interpreting the Benchmark Report

As with any data, caution must be exercised in making inferences from the benchmark report. It is important to frame any interpretation within the context of the school's environment. Consideration of external information about the Science curriculum, instructional practices and data from other classroom assessments is critical to making appropriate and meaningful inferences from this report. Interpretation of this report should also take the following factors into account:

- The number of items on each report corresponds to the number of items on the assessment, as outlined in the test specifications of each grade. This feature is specific to the Science MCA.
- There may be more than one item assessing a particular benchmark.

There are several misinterpretations that should be avoided:

- Color/shape and position of markers in the graphs **do not** reflect benchmark difficulty.
- Color/shape and position of markers in the graphs **do not** correspond to achievement levels (i.e., Does Not Meet, Partially Meets, Meets, or Exceeds the Standards).
- When comparing Benchmark Report graphs from different schools within a district, be aware that the range of values on the horizontal axis CLES scale is adjusted to fit each school's data. If a school has a large outlier (i.e., a benchmark with very high or very low relative performance) the graph will have a greater range reflected on the horizontal axis, and its benchmark markers will appear to be clustered more tightly together than those for a school with a smaller range of benchmark CLES values.

The primary purpose of the MDE Benchmark Report is to provide information to help curriculum and instructional staff in making inferences about their instructional/curricular activities and their students' level of understanding, based on performance data from the online Science MCA-III. The purpose of data in this report is **not** to designate strengths and weaknesses in the school. Rather, the Benchmark Report is to serve as a guidance tool to identify possible gaps in instructional content that the school staff may find relevant and important. In particular, it is important to recognize that this report reflects data on a sample of student testing behavior obtained at a single time point in the academic year, and may not fully reflect the systematic instructional and curricular outcomes as a whole. Furthermore, some of the results may depend upon the timing and sequence of when content was presented during the school year. For those reasons, it is critical to appropriately involve knowledgeable instructional staff in the discussion and interpretation of the results, and in deliberations about their implications for curriculum and instructional activities.

Technical Details for the Science MCA-III Benchmark Reports

Relative Benchmark-Item Performance and Common Language Effect Size

The relative performance on benchmarks items for students within a school or district is reported using the Common Language Effect Size (CLES). The CLES is a non-parametric statistic used to summarize group differences. The basic notion is that two groups (say, Group A and Group B) exist, where each group member has a score on an outcome of interest. The CLES is calculated as the probability that a randomly selected member from one group (e.g., Group A) will have a higher score than that of a randomly selected member of the other group (Group B). When the group score distributions are equivalent, the probability will be 0.50. As scores

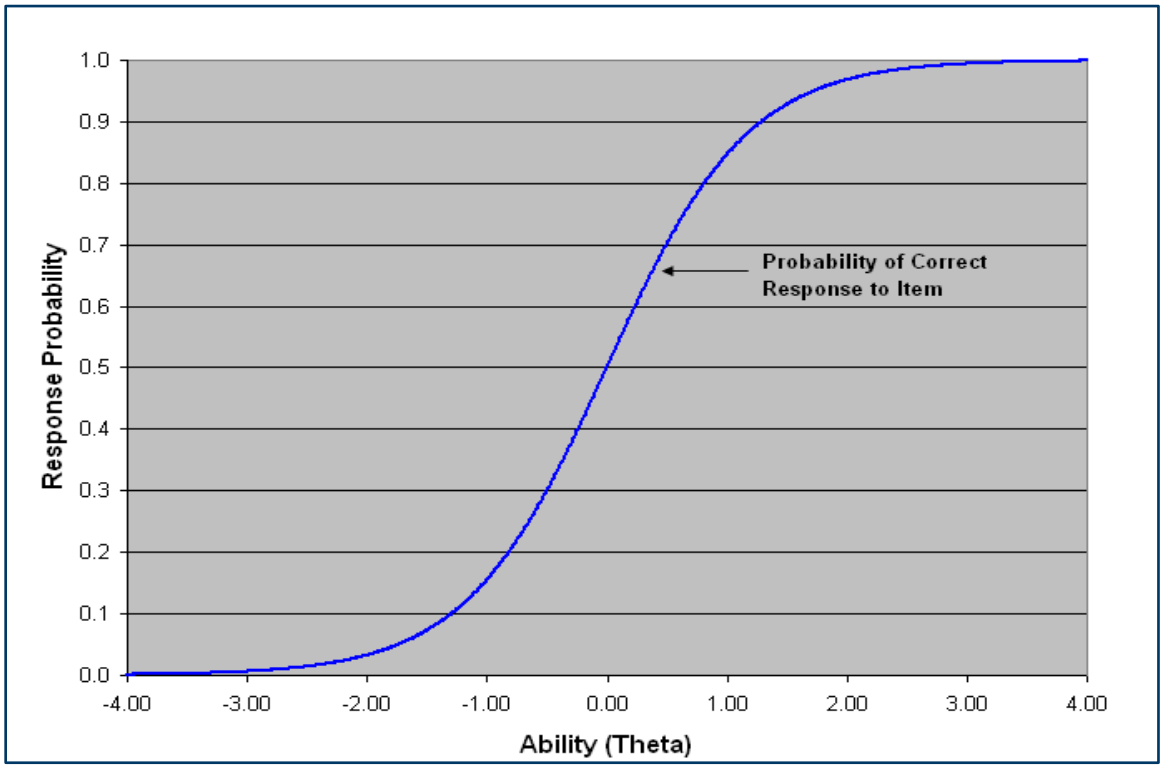
in Group A become increasingly higher than those in Group B, the probability that the score of a randomly selected member from Group A will be greater than that of a randomly selected member from Group B increases correspondingly, and the CLES becomes increasingly greater than 0.50. Conversely, as scores in Group A become progressively lower than those in Group B, the CLES will move progressively lower than 0.50.

The Science MCA-III Benchmark report uses the CLES to compare performance of two groups on each item administered. The scores comprising the first group are the observed item scores for students in the school. The scores comprising the second group are the expected item scores from each student, given their overall score on the Science MCA-III. These expected scores (or average conditional performance) are calculated based on the 3-parameter logistic (3PL) measurement model that underlies all scaling on the MCA-III. The 3PL model describes the probability of a correct response on each benchmark item, given a student's overall MCA-III score (see Fig. 2).

The School-Based CLES expresses the probability that a student selected at random from the school received a higher score on an item than a student drawn at random from a group whose overall MCA-III score distribution was identical to that of the school, but whose item scores were those expected based on the 3PL model (i.e., average conditional performance). When observed item response scores are equivalent to those expected based on overall MCA-III scores, the result will be a CLES of 0.50.

For each item, the average conditional performance for a school/district is represented as 0.50 on the School-Based CLES scale, which is interpreted as a student drawn from a random conditional distribution for the school/district having a 50/50 chance of exceeding the expected performance of another student drawn from the same random conditional distribution on any particular item.

Figure 2. Sample item response function: Probability of correct response conditional on ability.



CLES values greater than 0.50 indicate that the observed student performance on the item at the school exceeds the expected conditional student performance. CLES values less than 0.50 have the opposite implication: school performance is lower than expectation given the ability of students that were administered that item. When CLES is calculated based on dichotomous item scores, the deviation from CLES value 0.50 is approximately equal to one-half the difference in proportions correct in the two groups. Thus, a benchmark CLES value of .55 can be interpreted to mean that the observed proportion of correct responses to benchmark items is 0.10 greater than the expected proportion correct.

School or District Reference Line

Within the graph for each strand, a gray dashed vertical line at the 0.50 position on the horizontal axis represents expected school or district performance for each item. This expected performance is based on the total ability score of students within the district/school and effectively anchors each item at CLES = 0.50. Thus, the gray dashed vertical line reflects performance on an item right at expectations based on the total test score for all students administered items.

State Reference Line

Although the focus of the Benchmark Report is within-school comparisons of observed and expected benchmark item score distributions, some users may be interested in comparing school and statewide performance on the CLES scale. The heuristic approach adopted in the benchmark report is to calculate the expected count of correct responses if a student whose science ability was at the state average for the grade was administered the same items actually taken by students in the school. As before, a CLES index is calculated, this time comparing the expected correct response count (across all items) for the average state student vs. the students in the school. The gray solid vertical line plotted at the obtained CLES value represents state-wide performance relative to the school or district. When the solid state reference line is to the right of the dashed school/district line (i.e., >0.50), it means the expected overall state performance exceeded that of the school. Conversely, when the state reference line is less than 0.50, it indicates expected state overall performance that is lower than expected performance for the school.

95% Credible Interval Bands

The credible interval is the Bayesian analogue of the confidence interval reported in common statistical (i.e., frequentist) practice. The 95% credible interval band reported here can be interpreted as the range of CLES values within which there is a 95% probability the true CLES value lies, given the observed data. The 95% credible interval bands are estimated empirically, based on observed CLES values resulting from 20,000 paired random draws from beta distributions with parameters (1+observed number correct, 1+observed number wrong) for the school observed data and (1+expected number correct, 1+expected number wrong) for the school expectation data. The .025 and .975 quantiles of the observed CLES sampling distribution serve as the limits of the 95% credible interval band. One consequence of this empirical approach is that when a district has a single school at a grade, the sample distributions from school and district analyses can differ very slightly, and the resulting school and district benchmark graphs will be not quite identical.