Georgia Department of Education

Educating Georgia's Future

Georgia Milestones Spring 2021 EOG and EOC Scores
Summary of Evaluation and Score Interpretation Guidance

The results of the Georgia Milestones Spring End-of-Grade (EOG) and End-of-Course (EOC) administration follow extensive learning disruptions due to the COVID-19 pandemic. Critical aspects of Georgia Milestones remain consistent despite these disruptions, including: the academic content standards, the achievement standards, the administration format, and the scoring procedures and data-quality criteria. However, some key factors have changed, which necessitates caution and context when interpreting individual or summary scores: many students received virtual instruction following interruptions and closures, opportunity to learn has been variably reduced due to health and safety measures in the past year, the contribution of EOC scores to final course grades has been reduced in weight as a logical precaution and fewer students participated in this Spring administration as compared to prior years. Before, during, and following the administration of Georgia Milestones 2021 Spring assessments, standard operational analyses and quality assurance analyses were completed to evaluate and ensure the stability and accuracy of reported scores. This brief summarizes those analyses and results and provides additional context for stakeholders using these scores in decision making.

Overall, these results meet the rigorous reliability standards of the Georgia Milestones assessment program and are valid when interpreted in context: as one measure of a student's achievement towards mastery of the state's academic content standards in the face of unprecedented challenges.

Before: Planning the Analysis

Prior to the Spring administration, psychometric plans for equating to a common scale for score comparability were evaluated in detail and approved by Georgia's Assessment Technical Advisory Committee (TAC). This calibration and equating method uses item response theory (IRT). IRT has been widely adopted for test score scale maintenance across the assessment industry, including essentially all states. Preliminary research was completed to evaluate expected classification consistency under the updated test design and domain reliability under pre-equated conditions. In evaluating this research and the recommendations of the TAC, the decision was made to use pre-equated parameters where possible, to ensure stability in the solution, and post-equate only a few new items where necessary, pending data review. In a typical year, over 95% of Georgia Milestones tests are fully pre-equated, a process which uses information from previous administrations to ensure that scores from different versions (years and forms) of the assessments are comparable. In the cases where a post-equated model is used, around 97% of the items remain anchored to their pre-equated parameter, and a few which demonstrate drift are modeled with their post-equated parameter, using data from the current administration. This year, the recommendations from our TAC, the Council of Chief State School Officers, the National Center for Improvement of Educational Assessment, and other experts in the field all were to use the pre-equated item-parameters where possible, as this solution is based on stable data from Georgia students under normal learning conditions, and will support (with the context and cautions below), longitudinal comparisons and scale stability.

During: Evaluating the Results

Several cycles of standard operational psychometric analyses were completed and supplemented by additional quality-control steps designed to identify and mitigate any potential instability from this year's learning disruptions. The four primary considerations for evaluating the results for each examination were: reliability, data-model fit, representative sampling, and impact on achievement-level classification decisions. Total test reliability by form was a top consideration, and this was compared against rigorous reliability

criteria, as well as the reliability outcomes from prior administrations. This administration's results indicated an average reliability above .9 across content areas and forms. This level of reliability is considered excellent, and is comparable to historic reliability for this program. Data-model fit was evaluated, and any misfit was flagged using the same flagging criteria as in typical operational years. For all grade/content areas/courses, fit was excellent, with rates of misfit being at or below rates identified in prior years. Items with post-equated parameters were rigorously evaluated to ensure stability and quality, and all post-equated parameters included in the final scoring model were confirmed to be free of significant misfit and contribute positively to the precision and stability of overall scores. When evaluating the sample, in addition to ensuring the data were sufficient to produce stable estimates, the representativeness of the sample by gender, ethnicity/race, grade, and region/ RESA was closely monitored, and most groups were found to be within 5% of the distribution observed in prior years. This indicates mostly consistent representativeness, despite a reduction in the total sample, though some demographic differences were observed as compared to prior years in the area of region and ethnicity, with slightly lower representation from Metro areas. While all sample criteria were met to produce valid and reliable individual scores, this does necessitate caution when interpreting summary scores, such as summaries by school, district, and region. General guidance is offered in the next section on interpreting summary scores in the context of this year's learning disruptions, and further research is in progress on this point. The final consideration was the impact of including the few post-equated parameters calibrated with this administration's data on achievement level classification. Differences observed when including these parameters were consistent with expectations of improved classification accuracy. Student achievement classification results are considered to be statistically sound.

After: Using the Scores

While the results above do support the reliability and validity of Spring 2021 EOG and EOC scores, the following guidance should be considered when interpreting individual and summary scores from this administration:

Individual scale scores, achievement levels, domain scores, etc. should be interpreted as one measure of a student's mastery of the knowledge and skills outlined in Georgia's academic content standards. These scores are most meaningful when considered in the context of learning and any associated extenuating factors. For example, a student's performance may classify them as a Developing Learner, indicating the student mastered some, but not all, of the academic content standards. However, these scores cannot indicate whether the student had the opportunity to learn *all* of the content standards or whether, due to pandemic-related learning disruptions, the student only had the opportunity to learn *some* of the content standards.

Summaries of Georgia Milestones scores by class, school, district, and state should likewise be interpreted as one measure of mastery of the knowledge and skills outlined in the state's academic content standards. These scores should not be used as a part of a longitudinal trend analysis without including context of this year's pandemic and associated learning disruptions, and varying access to instruction. For example, changes in summarized performance could not be attributed to program or curricular choices. Any difference in outcomes as compared to prior years cannot be interpreted in isolation from the impact of pandemic-related disruptions to teaching and learning. Additionally, participation rates and representativeness across demographic subgroups should be considered, and extra caution should be taken when interpreting a summary of achievement which is comprised of a low percentage of enrolled students tested or comprised of an unrepresentative sample based on demographics or prior achievement. Further research is in progress on this point.

Overall, these results meet the rigorous reliability standards of the Georgia Milestones assessment program and are valid when interpreted in context: as one measure of a student's achievement towards mastery of the state's academic content standards in the face of unprecedented challenges.