GAA 2.0 Spring 2021

Summary of Evaluation and Score Interpretation Guidance

The results of the GAA 2.0 Spring 2021 administration follow extensive learning disruptions due to the COVID-19 pandemic. Critical aspects of GAA 2.0 remain consistent despite these disruptions, including: the alternate 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, and fewer students participated in this administration as compared to prior years. Before, during, and following the administration of the GAA 2.0, standard operational analyses and quality assurance analyses were completed to evaluate and ensure the stability, accuracy, and technical quality 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 GAA 2.0 and are valid when interpreted in context: as one measure of a student's achievement towards mastery of the state's alternate academic content standards in the face of unprecedented challenges.

Before: Planning the Analysis

Prior to the Spring 2021 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 virtually all states. In a typical year, the GAA 2.0 will be post-equated to bring new operational items to the base scale. 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.

Preliminary research was completed to evaluate expected precision and reliability if only pre-equated item parameters are used. 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 new operational items where necessary, pending item data review. Accordingly, this year's plan prioritized pre-equating all possible parameters, and includes some additional evaluation steps to ensure the stability of any included post-equated parameters.

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 three primary considerations for evaluating the results for each examination were: reliability, data-model fit, and representative sampling. Total test reliability by form was a primary consideration, and this was compared against rigorous reliability criteria, as well as the reliability outcomes from prior administrations. This administration's results indicated comparable reliability to prior results for this

program, with a range of .8 to .9 across content areas and grades. This level of reliability is good, and is comparable to historic reliability for this program. Data-model fit were all evaluated, and any misfit was flagged using the same flagging criteria as in typical operational years. For all grade/content areas, fit was excellent, with rates of misfit being at or below rates identified in prior years. Data from 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, and grade was closely monitored. Among the students receiving instruction on Georgia's alternate academic content standards who enrolled in the 2021 administration, around 72% took the test. Sample representativeness checks confirmed that most groups were found to be within 5% of the distribution observed in prior years, thus ensuring consistent representativeness despite a reduction in the percent of enrolled students tested. This indicates mostly consistent representativeness, 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. To ensure summary scores from this year are appropriate for interpretation, we further examined the prior achievement of the students who tested this to identify, if present, any differences indicating the students who did test this year are not representative of the total population. This analysis investigated the prior achievement of the students who tested this year to identify any significant difference from the prior achievement of the total enrolled population. Results of this analysis indicated that while the percent of enrolled students tested this year was lower than typical years, the students who completed the GAA 2.0 tests in 2021 are meaningfully representative of the full enrolled population based on their prior achievement. This research validates the use and interpretation of achievement summaries from this year, with the caution outlined below.

After: Using the Scores

While the results above do support the reliability and validity of GAA 2.0 scores, the following guidance should be considered when interpreting individual and summary scores from this administration:

Individual scale scores and achievement levels should be interpreted as one measure of a student's mastery of the knowledge and skills outlined in Georgia's alternate 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 Level 2, indicating the student mastered some, but not all, of the alternate 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 GAA 2.0 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 alternate 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.

Overall, these results meet the rigorous reliability standards of the GAA 2.0 assessment program and are valid when interpreted in context: as one measure of a student's achievement towards mastery of the state's alternate academic content standards in the face of unprecedented challenges.