Measuring the Power of Learning.®

Georgia Assessments for the Certification of Educators® (GACE®) Validity Argument

Spring 2018

TEST DESIGN

The purpose of the Georgia Assessments for the Certification of Educators® (GACE®) is to help the Georgia Professional Standards Commission (GaPSC) ensure that candidates are ready to perform the job of an educator in Georgia's public schools. They are designed to measure the knowledge and skills relevant and important for entry-level teachers. The assessments are taken by individuals seeking admission to an educator preparation program and/or seeking educator certification in the State of Georgia. Candidates who are enrolled in or seeking admission to a Georgia state-approved educator preparation program must be approved to test by their program providers in order to take a GACE assessment for Georgia certification. Otherwise, they will need to seek approval from the GaPSC.

The GACE assessments indicate a candidate's readiness to begin practice (or an approved teacher education program); they are not intended to predict the candidate's future performance as an educator, advanced professional, paraprofessional, or a teacher education student. In general, the claims for GACE assessments are as follows.

- Candidates who pass GACE content assessments have demonstrated sufficient knowledge and skill to begin practicing.
- Program Admission assessment candidates who pass have demonstrated enough knowledge and skills to enter into an educator preparation program.
- The certificate upgrade assessments are for candidates who are already certified educators. Candidates who pass these assessments have demonstrated enough knowledge and skills to be an advanced professional in his or her field, and to provide services relevant to the certificate sought.
- Candidates who complete the Program Entry level of the Georgia Ethics assessments have had the opportunity to familiarize themselves with the Code of Ethics for Educators and with the principles of ethical decision making in an educational context. Candidates who complete the Program Exit level, have demonstrated an understanding of the Code of Ethics for Educators and an understanding of how to apply ethical decision-making principles in an educational context.

To support the claims of the assessments, all GACE assessments are aligned with the Georgia standards for the P–12 curriculum and with state and national content standards. The content specifications for each GACE assessment were developed in partnership with a diverse group of Georgia educators from both P–12 classrooms and educator preparation programs, and ongoing development involves Georgia educators throughout the process.

We collaborate with these educators at various times in order to learn their perspectives (e.g., the critical knowledge and skills for beginning practice that need to be measured, the candidates who will take the assessments). The evidence collected at these stages supports the content validity of the assessment and the claim that a candidate who passes the assessment has demonstrated the requisite knowledge and skills to begin practice.

TEST DEVELOPMENT

We follow an evidence-centered design (ECD) process for developing teacher licensure assessments. This process begins with a principled evaluation of the domain of interest by both assessment experts and experts in the field being assessed, and ends with an assessment that is aligned with relevant standards and that measures the content deemed important by experts for entry into the profession. The process ensures that we focus on the evidence required to be confident that a candidate has the appropriate — and appropriate amount of — knowledge and skill for beginning practice. The process is outlined below.

For every new assessment, ETS assessment specialists first create draft content specifications based on state and, where appropriate, national standards in the field being assessed. The specifications undergo a rigorous review process, involving senior assessment development and editorial staff, to help ensure that the draft is clear, concise, and consistent with ETS standards. Once these internal reviews are completed, the specifications are ready for review by a Georgia advisory panel.

Each GACE assessment is developed in partnership with an advisory panel consisting of 12 to 15 Georgia educators, including both teachers of the content area and faculty who teach the content area in educator preparation programs. Advisory panel members are chosen by the GaPSC to represent the diversity of educators in the state. The panel works closely with ETS assessment specialists to define the content specifications and confirm their alignment to the relevant standards. Throughout the process, panelists focus on confirming that the content measured by the assessment is both important and necessary at the time of entry into the profession. This process helps to build the content-related validity evidence necessary for licensure use.

After the test content specifications are finalized and approved by the advisory panel, the test development process begins. ETS recruits and trains selected educators to generate items to match the content specifications. We provide item writers with a scaffolded training process that facilitates the development of test items that align with the specifications (and, thus, the standards on which they are based) and meet standards for technical quality. Item writers are trained in the *ETS Standards for Quality and Fairness*, which includes explicit training in how to ensure that all test items are free from bias and are fair to all test takers. The training covers different types of test items (selected-response, constructed-response, etc.) as appropriate, and it includes discussions of sample items at different stages of the review process. After the training, item writers continue to receive feedback on their work from assessment specialists, so that the test materials they create are fair and measure the standards they are intended to measure.

Once the pool of questions has been developed and reviewed, those items are used to build test forms, which align with the test specifications. Each form follows a rigorous assembly and review process and receives multiple reviews, including a review by a standing committee of Georgia educators.[1] Similar to the initial advisory panel, the standing committee consists of P–12 educators and faculty from educator preparation programs. This group reviews each test form to ensure that it aligns to the content

[1] GACE Ethics does not follow the same form creation process as the other GACE assessments. Once Ethics test items are approved, they go into an operational pool. The delivery system then generates item sets randomly from the pool, creating unique forms for each candidate.

specifications, that it measures content important for beginning practice, and that the test questions are fair and valid for the purpose of the assessment.

STANDARD SETTING

To set the initial passing score for a new GACE assessment, ETS conducts a standard-setting study. When conducting standard-setting studies, we convene a panel of educators with expertise in the subject matter as well as in the mentoring or training of beginning teachers. These educators provide their expertise during the study. GACE panelists are selected on the basis of recommendations from the GaPSC. Invitations are extended to teachers of the content area, as well as faculty from educator preparation programs. Most are or have been certified in the state of Georgia for the certificate associated with the assessment.

The panel is trained to make probability-based judgments using a modified Angoff method when judging dichotomously scored (i.e., selected-response and numeric entry) items. When judging the constructed-response items, the panel is trained to use an Extended Angoff method. The judgments form the basis for a recommended passing score. Some GACE assessments are single, stand-alone tests, but in many content areas GACE assessments are made up of two or more tests (also referred to as "subtests"). The same standard-setting panel provides passing score recommendations for each subtest using the same process that is used for a single assessment. Each standard-setting panel makes two rounds of judgments with feedback and discussion between the rounds. The results of each round of judgments are included in the standard-setting technical report, and the panel's recommended passing score is provided to the GaPSC for a final decision.

ADMINISTRATION AND SCORING

Most GACE assessments are computer-delivered with selected-response questions or a combination of selected-response and constructed-response questions. The Assessment of Sign Communication-American Sign Language (ASC-ASL) is an interview-based assessment that is video recorded for later scoring. The Teacher Leadership assessment is a portfolio-based assessment, requiring candidates to submit written responses and artifacts (e.g., student papers). More information about the Teacher Leadership assessment and the ASC-ASL assessments is posted on the GACE website.

The accessibility of tests is considered in the design process as well as during test administration. ETS follows the *ETS Guidelines for Test Accessibility* and offers appropriate accommodations for candidates. Testing accommodations are available for test takers with disabilities or health-related needs who meet ETS requirements. All test takers requesting accommodations must register through ETS Disability Services. The information for test takers who need accommodations is available on the GACE website. Additionally, there is a supplement to the GACE *Registration Bulletin* that contains procedures for requesting testing accommodations and a registration form.

Test security is a serious concern. Before, during, and after the test administration, the testing centers ensure that candidates follow specific guidelines. Test takers are also subject to various security checks and forms of ID comparison to ensure the security of test content. Test takers who refuse to participate in these security measures, are not permitted to test and, as a result, forfeit their registration and test

fees. An overview of what candidates can expect on testing day is provided on the GACE website. Additionally, preparation materials describe the expectations of candidates during the testing period.

ETS's Office of Testing Integrity (OTI) employs approximately 46 investigators and supporting personnel, and coordinates assessment security services worldwide. These staff help keep assessments secure through the use of rigorously defined procedures and effective training of test administration personnel. The OTI specializes in three steps of test security: prevention, detection, and remediation.

Scoring begins immediately after testing windows close. For the selected-response items, a preliminary score is machine-calculated after the test is completed and the test taker is provided with this information. However, those preliminary scores are unofficial and cannot be used for any official purpose. The official scores are released to the test takers after the data from the test administration is completely analyzed and reviewed. For tests with constructed-response items, examinees may also receive analytic feedback regarding performance on the constructed-response section if the score received is sufficiently low.

ETS follows industry standard constructed-response guidelines when designing and scoring constructed-response items. When constructed-response items are scored, raters are trained to follow the rubric and maintain accuracy in its application. During scoring sessions, chief readers monitor the progress of constructed-response scoring. For example, previously scored papers are often included in a rater's folder in order to determine if that rater would assign the same score. The people who conduct the scoring are educators who are experts in the field assessed.

PSYCHOMETRICS

For each of the GACE subtests, the scaled scores are individually reported as well as the candidate's Passed or Not Passed status. Scaled scores are reported on a 100 – 300 range. The total raw test score is the sum of the correct dichotomously scored (i.e., selected-response and numeric entry) items and the total of the constructed-response items (after the appropriate weight has been applied). The constructed-response scores are assigned by two raters. The total raw scores are converted into scaled scores after the total raw test score is computed. Score users are informed that scaled scores are comparable against all versions of the subtest of an assessment but not across different subtests.

For most of the GACE content tests, tiered passing standards have been established by the GaPSC. Candidates can pass at the induction level (220 scaled score) or at the professional level (250 scaled score). A candidate must pass all subtests in an assessment in order to pass the assessment. All subtests must be passed at the professional level for the candidate to be considered passing the assessment at the professional level. If a candidate passes all subtests, but at least one is passed at the induction level, the candidate is considered to have passed the entire test at the induction level. The Program Admission assessment and the certificate upgrade assessments do not have tiered passing standards. For the Program Entry level of the Georgia Ethics assessments, candidates receive a percentage correct and a status of complete or not complete for each end-of-module test. Although they do not receive an overall test score, they do receive an overall status of "Completed" or "Not Completed" for the assessment. They must complete all training modules and end-of-module tests to receive credit. For the Program Exit level of the Georgia Ethics assessments, candidates receive a percentage correct and a status of

completed or not completed for each end-of-module test and the summative test. Although they do not receive an overall test score, they do receive an overall status of "Passed" or "Not Passed" for the assessment.

After the assessments are administered, statistical analyses are conducted to learn if the assessment is performing as intended. These data are collected to support the validity claims of the assessment. The claims are that the test is consistently and appropriately measuring the subject-matter knowledge and skills the candidates demonstrate.

Preliminary Item Analysis. We conduct Preliminary Item Analysis (PIA) on the selected-response items to evaluate test items' performance (i.e., item difficulty and discrimination). Based on item analyses, we identify items exhibiting questionable statistical characteristics. Also, the item analyses results are used to compare common item performance in the new form and reference form, when applicable.

If any evidence collected determines that changes are needed to maintain the test claims, immediate actions are taken. For example, if items are discovered to not be performing as designed or if an item is no longer secure, it is not included in a candidate's score. Items that have not previously been scored can be used to substitute the newly non-scored items. If this happens on first administration of an assessment, the standard-setting data is recalculated. The scores are then reported based on adjusted standard-setting results. Otherwise, the problematic items are taken out, and the test form is equated with the new set of scored items.

Differential Item Functioning. To identify and evaluate items that unfairly favor one group of examinees compared to another, Differential Item Functioning (DIF) analyses are conducted. To meet DIF volume requirements, there must be at least 300 examinees in the focal group and at least 700 in the focal and reference groups combined. Focal groups are defined as the following: African American or Black, Hispanic, Asian, Native American, and Female. Reference groups are White and Male examinees, as relevant. Only examinees who indicate that English is their best language of communication, and who indicate that they first learned English, or English and another language, as a child are included in the analyses. Items identified to have DIF are reviewed by a DIF panel to determine if the differences in performance indicate actual item bias. Research of DIF analysis for constructed-response items is ongoing. Currently, DIF analysis is not required for constructed-response items. GACE Ethics assessment end-of-module and summative test forms contain different sets of items for each candidate that are randomly selected from the item pool; the volume on each form does not meet DIF analysis requirement. Therefore, a DIF-like group item performance analysis will be conducted to evaluate item performance differences in different gender and ethnicity groups by aggregating candidates' data on items across different testing occasions.

Equating.[2] For selected-response tests, equating of new forms and most revised forms is usually done by the common-item equating methods. A choice of a particular equating method depends on factors such as sample size available during equating, relationship of the new and old form scores (i.e., linear or

[2] Does not apply to Georgia Ethics assessments

non-linear), ability difference between the new and old form samples, etc. For certain small-volume selected-response tests, the single group nearly equivalent test (SiGNET) design equating is used where two forms are embedded in one form with most of the operational items overlapping and a small set of pretest items designated as operational items on the second form. This way, data can be accumulated on the first form to equate the second form using the single group equating design. The raw-to-scale conversion for the reference form is then applied to the raw-to-raw conversion to get the raw-to-scale conversion for the new form.

After conversion tables are produced, but before scores are released, a graph depicting means by administration month (trend plots) and a table of descriptive statistics are produced for tests with scaled score means out-of-range for the current administration compared to previous administrations. Final item analysis (FIA) is also conducted to evaluate test item as well as form statistical characteristics. A technical report is written for each test describing its psychometric characteristic and scaling information.

Reliability. Test reliability refers to the likelihood of a candidate obtaining the same score on a different test form or at a different occasion. We compute both test-level internal-consistency reliability of the reported scores and several indices of rater reliability, including rater agreement and the intra-class correlation between scores for the constructed-response items. We generally use Coefficient Alpha as the index for estimating the total test reliability for a single form of a test consisting of a single construct. Any test with lower than desirable reliability is reviewed, and recommendations are provided for future test assembly. The GACE Ethics assessment test form taken by each candidate is individually generated by random item selection from the item pool. That is, test forms are most likely different for each candidate, making it impossible to calculate internal-consistency reliability for each form. Therefore, only part to whole correlations between end-of-module scores and summative scores are calculated to provide some convergent evidence of how consist test scores can be between different parts of the tests with the summative test.