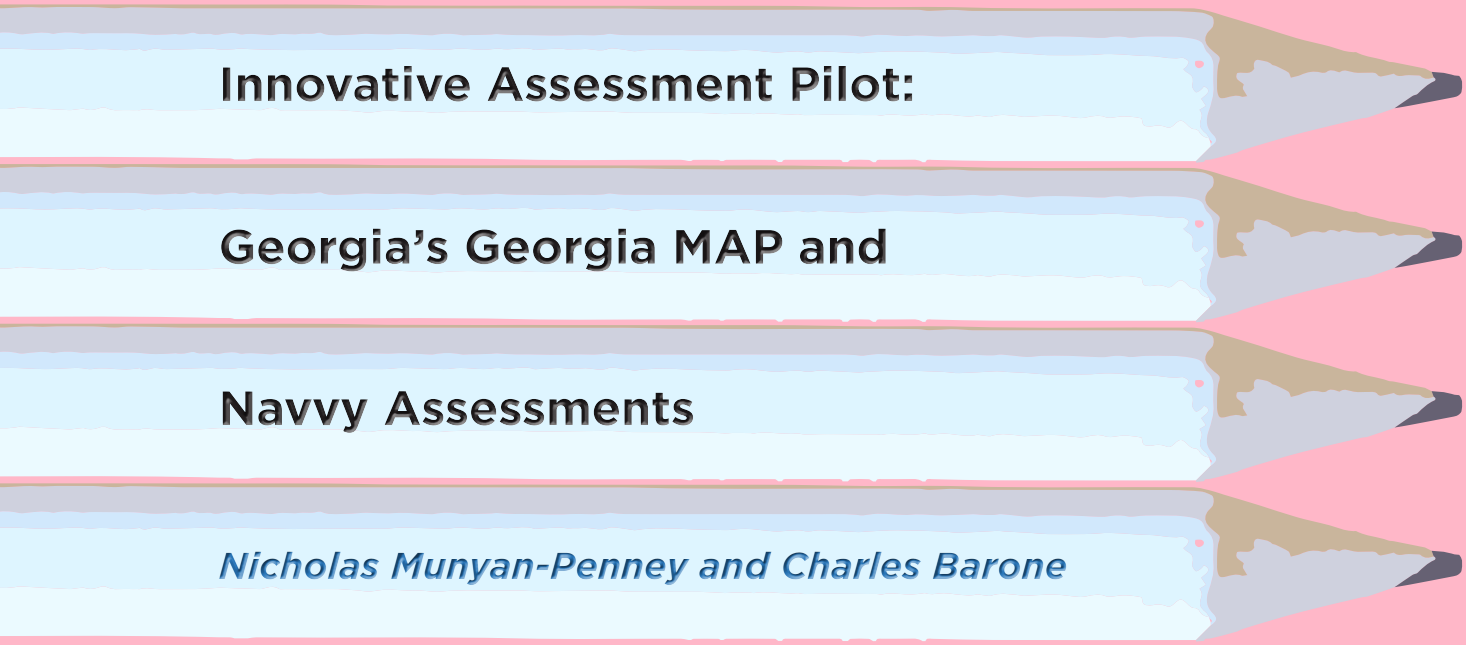

Innovative Assessment Pilot: Georgia's Georgia MAP and Navvy Assessments

Nicholas Munyan-Penney and Charles Barone

DECEMBER 2020

ERN

EXECUTIVE SUMMARY

Georgia is one of the first four states to be approved to participate in the Innovative Assessment Demonstration Authority (IADA) under the Every Student Succeeds Act (ESSA). Districts participating in the state's pilot are free of federal requirements that the same summative assessments be administered in math and English Language Arts (ELA) in grades 3-8 and that all students in the state, with some exceptions[1], participate in the same statewide assessment.

Georgia's IADA pilot is motivated by two main goals: to reduce student testing time and to implement an assessment system that educators can use to inform instruction throughout the school year. However, rather than selecting a single innovative assessment which will be scaled over the course of the pilot, the Georgia Department of Education (GaDOE) decided to leverage existing innovation throughout the state with a competition-like process. After a thorough vetting, three districts/LEA consortiums were selected to participate in IADA, two of which were approved by the US Department of Education.

Georgia's IADA pilot is motivated by two main goals: to reduce student testing time and to implement an assessment system that educators can use to inform instruction throughout the school year.

Over the course of the five-year pilot, GaDOE, along with the State Board of Education (SBOE) will serve as project managers for the pilot, overseeing implementation, providing technical assistance, and working to integrate the assessments into the state's larger accountability system. In the final year of the pilot, GaDOE will contract with an external vendor who will evaluate each system, with a particular focus on alignment to the state's academic standards and comparability with the current assessment system, Georgia Milestones. Based on the results of these evaluations, GaDOE and SBOE can select one of the innovative assessments to scale statewide or decide to continue using Georgia Milestones. If the state selects one of the innovative assessments, GaDOE will apply for a two-year extension of IADA in order to scale the chosen assessment statewide.

Opportunities. Georgia's unique approach to IADA has some potential advantages:

- It allows GaDOE to harness existing innovation and select the assessment system that works best across the state, rather than going all in on a single system;
- Both assessments provide educators with actionable data throughout the year through interim reports that are directly related to what students are already learning in the classroom; and
- Georgia MAP's inclusion of above and below grade level test items could allow teachers to better address student learning gaps.

Risks. However, there are also a few inherent risks:

- Districts that have invested heavily in adopting one system may be resistant and/or lack capacity to adopt another system if theirs is not selected to be scaled statewide;
- Navvy's unique scoring system will create extra hurdles for GaDOE if they attempt to integrate it into their existing accountability system and establish comparability with Georgia Milestones;
- By allowing students to reassess, Navvy could be reducing reliability as reassessments could measure increased student familiarity with the assessment rather than academic growth; and
- While Georgia MAP's non-grade level items may improve instructional decisions it may limit the assessment system's overall validity.

OVERVIEW

Below are brief overviews of the two assessments being piloted:

Georgia MAP

A collection of schools currently using NWEA's formative MAP Growth assessments will be piloting a through-year assessment system. NWEA will lead the development of Georgia MAP, which will adapt MAP Growth to fully cover the breadth and depth of Georgia's state standards and assess growth and proficiency simultaneously. Georgia MAP will be administered three times over the course of the school year: in fall, winter, and spring. The fall and winter administrations will consist of adaptive, online assessments, and the spring administration will also include a writing-based summative performance task.

Interim reports will provide information on student growth and proficiency compared to national norms as well as state standards. In addition to providing information based on grade-level expectations, interim reports will provide teachers with data related to below- and above-grade expectations to help fill in knowledge gaps or appropriately challenge students. Georgia MAP will cover math, ELA, and science in grades 3-8, though NWEA has expressed willingness to expand the assessment into both high school and grades K-2.

CTLS-ASSESS: ASSESSMENT SYSTEM REJECTED BY ED

Interestingly, the US Department of Education (ED) opted to accept Georgia's overall assessment pilot program, while rejecting a single assessment system—CTLS-Assess—from the statewide competition. This move by ED is a welcome surprise, given Secretary DeVos' largely hands-off approach to K-12 accountability under ESSA. In its letter to GaDOE, ED cited concerns that there were no clear safeguards under CTLS-Assess to ensure the test fully measured student achievement in relation to state standards nor processes in place to determine reliability and comparability, both overall and for student subgroups.[2] We had similar misgivings during our review of GaDOE's IADA application.

In order to establish evidence of test validity, Cobb County School District (CCSD) planned to administer both CTLS-Asses and Georgia Milestones to a sample of students annually, but CCSD's student demographics aren't representative of the state, making these comparisons useless for informing scaling efforts. However, comparability across students and classrooms within the CTLS-Assess system was an even larger concern than that between systems. It appeared that teachers were able to design their own assessments, as well as use outside assessments within the system simply by creating answer keys. These assessments were supposed to be validated by assessment and curriculum leaders in the district, but it's unclear how rigorous this review process would have been. Additionally, constructed responses would have been scored by students' own teachers. While teachers would have received training on how to score items, there was no evidence that these items were to be double scored or vetted in any other way.

CTLS-Assess may function well as a formative assessment, but its inclusion in Georgia's accountability system would have raised serious red flags about both the future of Georgia assessments and, more importantly, ED's vetting process for IADA applicants. While a certain amount of risk is inherent in developing new assessments, as new states are encouraged to apply, it's encouraging that ED seems committed to at least ensuring minimum safeguards around validity and reliability are in place.

Navvy

The Putnam Consortium consists of a group of LEAs across the state of Georgia that are implementing Navvy (pronounced like savvy), a through-year assessment system designed to provide real-time data on student competency. Navvy was designed in partnership with Georgia educators specifically to measure student achievement against the state's academic standards. Assessments allow for flexible administration across the school year, and students have multiple opportunities to demonstrate competency in standards. Unlike most assessments, Navvy doesn't produce a raw numerical score, instead it shows binary results for each standard (competent or not competent), which

can be used to personalize instruction to individual students or small groups. Annual summative scores will be based on the percentage of competencies a student has mastered by the end of the school year and will not require a year-end summative assessment. Navvy will cover math, ELA, and science in grades 3-8 and high school.

VALIDITY AND RELIABILITY

In year five of IADA, GaDOE will commission an outside evaluation of each assessment system that will focus largely on criterion validity defined as comparability with the current assessment system, Georgia Milestones. Additionally, each individual pilot will be conducting their own analyses throughout the process:

Georgia MAP

NWEA has a robust strategy for ensuring comparability with Georgia Milestones. First, they will be working with local educators to ensure that assessments are aligned and cover the full breadth and depth of the state's standards. In year three of the pilot, the first year of the new through-year exam, all students will take both assessments. In following years, the Georgia Milestones items will be embedded in the Georgia MAP test in order to conduct linking studies.

Studies comparing the existing MAP Growth and Georgia Milestones scores already show strong correlations (r=0.79-0.87), and NWEA plans to conduct analyses and make revisions yearly to improve the assessments each year of the pilot. Georgia MAP will also include writing-based performance tasks, but to ensure internal validity, the scoring of these items will not be done locally. However, teachers will be trained on scoring

WHY COMPARABILITY MATTERS

One key theme throughout ESSA is that standards and assessments must be the same, statewide, for all students. The words "all," "same," and "statewide," as applied to standards, assessments, schools, and students appear, consistently, multiple times across what is really the heart of the entire 400-page law. As words go, "all," "same," and "statewide" are about as precise as it gets and, generally speaking, these provisions have a 25 year history under various iterations of the Elementary and Secondary Education Act.

Local assessment systems have broad policy and political appeal but two key reasons ESSA, outside the innovation pilot, requires that state assessments be the same for all students are:

1) Such measures cannot be compared against one another; and

2) Students in different local education agencies could be held to very different standards, even though they would ultimately be applying to the same colleges and competing for the same jobs.

Despite the best intentions, there are immense political and economic pressures at the local level to cast schools in the best light possible. If we abandon statewide assessment systems, poor and minority students, students with disabilities, and English Learners—who historically, prior to advent of the standards and assessment movement, were held to lower standards—might return to a time when they repeatedly were told they were doing fine, only to graduate from high school and discover they didn't have the skills needed to succeed in college and the workplace. Moreover, resources that are now allocated on the basis of accountability systems geared to a single and apples-to-apples comparable set of state tests—those, for example, for after-school and summer programs, tutoring, teacher training, and new curricula—might be misdirected away from areas that actually need them most, because each district or school would then be measured by different standards and different yardsticks.

While innovation is to be valued and encouraged, we need to be mindful of the reasons that statewide standards and assessment systems were implemented in the first place lest, in years ahead, we see an ever-accelerating race to the bottom.

procedures so they can use the results of the assessments to inform instruction as well as administer and score formative performance tasks throughout the year. Importantly, NWEA notes that while annual summative reports will be designed to be comparable to Georgia Milestones, interim reports produced from the fall and winter administrations will not comparable, but instead should only be used to inform local instruction.

Navvy

As with Georgia MAP, Navvy will evaluate the assessment annually for comparability by having a sample of students take both Navvy and Georgia Milestones and by embedding Georgia Milestones items within their assessments. However, taking a page from New Hampshire, Navvy notes that they are not focused on developing very strong score comparability to Georgia Milestones, as attempting to tell the same story about student achievement stifles innovation. Given this, it's not surprising that the greatest source of uncertainty and perhaps the greatest cause for concern is how the Navvy assessment system will create a summative score based on binary statements of competence and how these will correspond to current achievement level descriptors. Navvy plans to enlist teachers from participating schools to weigh in on what percent of standards mastered (or percentages of standards mastered of different difficulties) equate to each level of achievement which introduces no small amount of subjectivity into the process. Another concern for validity and comparability is students' ability to reassess throughout the year in order to prove competence on various standards. While pedagogically sound and in keeping with the vision of the Navvy system, this is a very different approach to assessment than Georgia Milestones which could artificially inflate students' scores based on familiarity rather than actual knowledge or skill.

Since both of the assessment systems involve not only a significant change in test administration, but also a fundamentally new approach to instruction by using assessment data, implementing any of these systems with fidelity in two years may be unrealistic.

Accessibility for students with disabilities and English learners

NWEA notes that all of its test items are designed using Universal Design for Learning (UDL)—design principles focused on ensuring maximum accessibility for all students, particularly SWD and EL. Additionally, NWEA had adopted language recommended by Council of Chief State School Officers (CCSSO)'s Accessibility Manual, and all items will be reviewed by a Bias, Sensitivity, and Fairness panel. Navvy also states that it uses principles of UDL in the development and review of all test items.

IMPLEMENTATION

As noted in the introduction, GaDOE has no intention of scaling either of the two innovative assessment pilots statewide by the end of the five years of IADA. Instead the state plans to select an assessment system in the final year based on its internal and external evaluations. If the state chooses one of the two innovative assessments, it plans a rather ambitious two-year scaling period, with one planning year before scaling the innovative assessment statewide.

Since both of the assessment systems involve not only a significant change in test administration, but also a fundamentally new approach to instruction by using assessment data, implementing any of these systems with fidelity in two years may be unrealistic. However, the number of LEAs who will be introduced to a new system in this scaling period will depend largely on the extent to which the assessments spread across the state over the course of IADA. Navvy, for instance, states that it could be used in as many

as 70% of Georgia schools by year five, so scaling this system could be significantly less arduous than Georgia MAP.

Any statewide scaling effort will be aided by the collective scaling of the two assessments over five years, which would give participating educators experience with through-year assessments. Additionally, both Georgia MAP and Navvy would be able to ease the state's professional development efforts through the deployment of external resources.

ENDNOTES

1 ESSA allows an alternate assessment for students with the most significant cognitive disabilities. The law and accompanying regulations cap the use of alternate assessments at 1% of all students statewide although a number of states have applied for and received waivers of the 1% cap.

2 https://www2.ed.gov/admins/lead/account/iada/gaiadaapproval2019.pdf

APPENDIX A

Reliability

Type	Explanation	Georgia MAP	Navvy
Reliability of student performance	One student should be able to take a test on Monday and then again on Tuesday and get very similar results each day.	Can use the through-year administration to establish this type of reliability.	Can use the through-year administration to establish this type of reliability.
Inter-rater reliability	The results should be the same no matter who scores the test and when they score it.	NWEA has extensive experience with scoring assessments but will need to focus more attention on ensuring the reliability of scoring on performance tasks that still being developed.	No items will be scored locally, and Navvy has extensive experience with scoring standardized assessments.
Reliability between different forms of the same test	Different forms of a test have slightly different questions in a slightly different order. However, the content and difficulty level are the same and a student should perform similarly on both tests.	NWEA has committed to annually study reliability and incrementally improve each year.	Navvy has committed to annually study reliability and incrementally improve each year.

Validity

Type	Explanation	Georgia MAP	Navvy
Construct Validity	The adherence of a measure to existing theory and knowledge of the concept being measured.	MAP Growth is a nationally used test aligned to commonly used standards—it's being retrofit for local state context. It also assesses non-grade-level content, which could weaken its ability to test grade-level content for accountability.	Navvy is specifically designed align with Georgia's academic standards.
Content Validity	The extent to which the measurement covers all aspects of the concept being measured.	Unlike the other pilots, NWEA will be adapting a different assessment to fit Georgia's standards, so it will need to prove its adjustments fully cover these standards.	Navvy is likely the strongest in this area, as the test has been designed from the ground up with local educators to align with the state's standards.
Criterion Validity	The extent to which the result of a measure corresponds to other valid measures of the same concept.	NWEA will be conducting studies to establish validity.	Navvy's unique binary scoring system will require additional effort to establish comparability with other tests.