GEORGIA INNOVATIVE ASSESSMENT PILOT PROGRAM

TECHNICAL ASSISTANCE ANNUAL REPORT

Matthew Gaertner | Markie McNeilly
Assessment Research & Innovation @ WestEd
July 2021

TABLE OF CONTENTS

GEORGIA INNOVATIVE ASSESSMENT PILOT PROGRAM

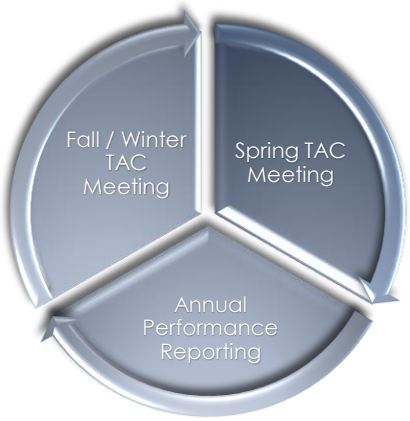
TECHNICAL ASSISTANCE ANNUAL REPORT

INTRODUCTION

When it was introduced as a provision of the Every Student Succeeds Act of 2015 (ESSA), the Innovative Assessment Demonstration Authority (IADA) was billed as a step toward decentralizing accountability and delegating to the states new choices about monitoring and incentivizing student learning. IADA was also envisioned as a force for assessment innovation, encouraging creative, flexible, and instructionally relevant testing programs that would bear slight resemblance to the standardized tests of today. The U.S. Department of Education has thus far awarded five states the authority to take up this challenge. Among them, Georgia is unique. Here—and not yet anywhere else—the IADA has seeded an intrastate competition, inviting multiple vendors to contend for a substantial prize: a statewide summative testing contract.

Two groups of school districts—the Putnam County Consortium (Putnam) and the GMAP Consortium (GMAP)—were granted the authority to develop new accountability assessments from the ground up, alongside each other. Over the course of a five-year pilot period, both consortia will have the opportunity to demonstrate that theirs is the assessment system suitable for adoption across the state. To support Putnam and GMAP, the Georgia Department of Education (GaDOE) contracted WestEd to provide technical assistance to both consortia and thereby advance what is now known as the Georgia Innovative Assessment Pilot Program (IAPP). The program's fall 2019 launch generated great interest from states, test developers, and researchers eager to watch two competing assessment ideas evolve together. Then, in Georgia and everywhere else, the 2019–2020 school year did not go as planned. Nevertheless, despite the challenges brought on by the COVID-19 pandemic, both consortia made progress—starting in earnest the work of building and stress-testing their innovative assessments. This report summarizes the activities, the accomplishments, and also the plans put on hold in 2019–2020 under the Georgia IAPP. The psychometric issues highlighted in the narrative are described in greater depth in the Appendices, which includes four Technical Advisory Committee (TAC) feedback reports—one for each consortium following two TAC meetings in winter and summer 2020.

PROGRAM REQUIREMENTS AND TECHNICAL ASSISTANCE PRIORITIES



This pilot program was authorized under Georgia Senate Bill 362 and by the United States Department of Education Innovative Assessment Demonstration Authority. Districts participating in the Georgia MAP Partnership and the Putnam County Consortium can administer a new assessment program (either the Georgia MAP Assessment in the GMAP consortium or the Navvy system of diagnostic assessments in Putnam) in lieu of the state's summative test Georgia Milestones, once they have demonstrated comparability between GMAP / Navvy and Georgia Milestones and received approval from the state.

THE COMPARABILITY REQUIREMENT AND THE THROUGH-YEAR APPROACH

In order to administer their assessments in lieu of Georgia Milestones, each assessment system will need to demonstrate comparability with the state assessment system and gain approval by GaDOE. This requirement is top of mind for most states participating in IADA, not just Georgia. However, the challenges presented by the competition-based format in Georgia are unique, and worth enumerating here.

First, it bears noting that the comparability standards imposed by ED do not appear overly stringent. Classification consistency at the performance level will likely suffice in the judgment of ED. This is a comparatively lenient standard, by intent: the hope is that granting latitude in comparability judgments will encourage pilot participants to begin implementation. In keeping with that spirit, GaDOE has not imposed unreasonable comparability criteria of its own or expressed dissatisfaction with a performance-level comparability standard (the TAC has yet to endorse one set of comparability criteria over another set; that discussion is scheduled for summer 2021).

In fact, the comparability challenge in Georgia is less about statistical comparability and more about the secondary considerations that establishing statistical comparability would trigger. Put simply, if an innovative assessment is given in lieu of Georgia Milestones, it becomes a statewide accountability assessment, immediately subject to all of the other criteria that statewide accountability systems must meet, by law. So, although comparability as a purely statistical criterion is not an impossible standard, it is also not the only standard. For example, test security throughout the testing window, appropriate accommodations, and evidence of fairness and reliability all become instant requirements after the relatively

simpler comparability bar is reached. For reference, we have included in this report the list of assurances these pilot programs agreed to provide before administration in lieu of Georgia Milestones (see the last page of the Appendices). The evidence and documentation required is extensive.

To complicate matters, the assessment systems competing for the statewide prize are through-year models. That means summative tests in September or October, which would require summer review from GaDOE. Furthermore, these systems will debut in the years following the COVID-19 pandemic, when school accountability designations (Comprehensive Support and Improvement / Targeted Support and Improvement) will be completely refreshed. Documentation of readiness to deliver an accountability test that can support those decisions will need to be thorough; a couple summer months will not suffice for key decision-makers review it. Specifically, each consortium will need to (1) develop the relevant documentation; (2) submit it to the TAC for feedback; (3) revise as needed; and then (4) submit it to GaDOE by the beginning of May to ensure careful review by September. If May arrives before spring testing ends under GMAP, Navvy, and Milestones, then it will be impossible for GMAP and Putnam to submit evidence of statistical comparability (e.g., performance level classification consistency for the students who take Milestones and one of the innovative assessments in the same spring).

To illustrate, let us consider the GMAP case. This consortium plans to finish field testing in 2021-22. In 2022-23, GMAP will administer operational assessments throughout the year and collect comparability data when Georgia Milestones scores arrive in June 2023. Suppose GMAP's final through-year assessment is also complete at that point and comparability analyses could be carried out instantly. Even then, it would be too late in the operational cycle for GaDOE to give GMAP a fair review. Approval would have to wait until the following year. **To wit, there is a strong possibility that neither GMAP nor Putnam will administer their assessments in schools for accountability purposes until 2024-25.**

TAC Meetings

One key source of technical guidance over the course of the IADA period is the Technical Advisory Committee, composed of nationally recognized experts in psychometrics and assessment policy, established in 2019 specifically for the Georgia IAPP. This TAC is a resource for Putnam and GMAP; it focuses specifically on the progress of their innovative assessments. The TAC convened twice in the 2019–2020 school year—once in December and again in June.

The December meeting focused on overviews of each assessment system along with feedback to support near-term objectives, most notably establishing comparability with

Georgia Milestones. Each consortium also discussed issues specific to their assessment systems. GMAP presented options for linking Georgia students' scores to the national RIT scale, which accompanies NWEA's MAP assessments. Putnam discussed options for covering Georgia's writing standards without a dedicated, stand-alone writing assessment administered every year.

Field testing for both consortia was impacted in Spring 2020 due to the pandemic, so the second TAC meeting focused more on annual performance reporting (due every summer to ED) and adjustments to proposed implementation plans as a result of COVID-19. Each consortium also began sharing plans for the logistical aspects of summative assessment. GMAP, for example, asked for feedback on its data security plans, while Putnam shared its plans for soliciting input from diverse stakeholder groups. While not particularly novel in and of themselves, data security and stakeholder input are basic elements of a testing program; innovations cannot get off the ground without them.

As a general rule, TAC feedback focuses on what the TAC members perceive to be "next up" in a nascent operational testing cycle. From the TAC's perspective, the most pressing issues at the outset were the logistical and administrative tasks (e.g., rostering) that account for a large portion of a statewide summative contract but that would be quite difficult for the small Putnam and GMAP teams to deliver without significant administrative scale up. In other words, the TAC was concerned that the pilot participants do not know what they do not know. Regular conversation about administrative and logistical issues has, however, resulted in fewer questions from the TAC. At present, the topic of primary concern is comparability. During the next meeting in summer 2021, the TAC, WestEd, and GaDOE intend to draft comparability guidelines for each consortium.

Timeliness of Materials for the TAC

In general, the TAC has found the introductions to both testing programs to be quite helpful, but TAC members had specific suggestions to help move both consortia along toward implementation. First, and most importantly, a TAC is only as helpful as it is prepared. Given the complexity of these assessment efforts, being prepared requires receiving documentation and materials in a timely manner. The GMAP Partnership has been careful to hand materials over either on time or close to it, however, Putnam has struggled. At each meeting thus far (since contract initiation) the TAC has received limited materials from Putnam without sufficient time to review them in advance of the all-day sessions.

The TAC has felt unable to provide Putnam with the in-depth feedback that will be required over the life of this contract. For the next TAC meeting WestEd is requiring materials be delivered in advance, and we will strictly enforce due dates, as we have previously. These

measures do not guarantee on-time delivery of materials to the TAC. However, we will introduce a new focus on reviewing TAC recommendations from the previous meeting and then discussing whether or not those issues have been resolved. It is reasonable to suspect that this step will encourage attentiveness and due date awareness.

Progress Toward Full Implementation

During each TAC meeting, each consortium presented their plan for implementing their assessment system. They posed questions to the TAC about establishing comparability between their assessment systems and Georgia Milestones, operationalizing their assessment systems, and how best to utilize data collected from their assessments through various studies. Both consortia had to push their timelines back in response to the COVID-19 pandemic, and although both were able to shift focus and continue to work toward implementation in schools, it is quite possible that neither assessment system will be approved for use in lieu of Milestones before 2024-2025 (see p. 5).

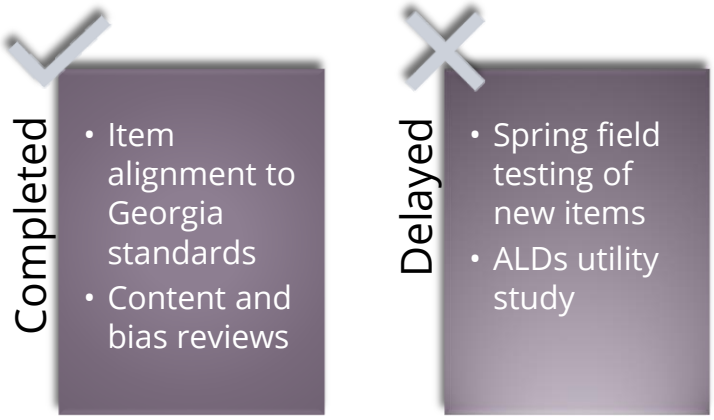
A Brief Note on Formative Assessments with Proposed Summative Uses

Through two TAC meetings, there has been little discussion of the potential advantages and dangers of using formative assessments to make summative judgments. Putting aside a debate about terminology let us assume that "formative" in this context includes interim, benchmark, and diagnostic tests. When those tests are reappropriated for summative use, what happens on administration day? How do attitudes toward the assessment change? Are there commensurate changes in score distributions? Do teachers grow more sophisticated in their data use? These questions tap largely non-psychometric elements of these programs' theories of change; it is nonetheless easy to imagine them emerging as the most vexing dilemmas introduced by the Georgia IAPP. In future TAC meetings, WestEd will encourage the consortia to grapple with the inherent tension between formative and summative test uses and consider how their guidance to their participating schools can ensure that, on the ground, their assessments are being used as intended and score interpretations are supported by validity evidence.

GMAP Partnership

At the December 2019 meeting, GMAP and NWEA shared their development plans for the 2020–2021 school year. Planned activities included beta testing their through-year system in the winter, then administering a stand-alone field test in the spring. Additionally, they hoped to run simulations for quality control on their developed adaptive tests. With content development already underway for ELA and Math, GMAP also planned to begin development of Science content.

During the 2019–2020 school year, the GMAP partnership and NWEA were able to proceed with many activities needed to scale up their assessment system. Educator committees were convened to review item alignment to the Georgia standards, content, and bias reviews. However, some activities had to be delayed. The largest setback was that the field testing that had been planned to take place in Spring 2020 will be pushed out to the 2020–2021 school year. Additionally, the second phase of their Achievement Level Descriptors utility study was postponed. This study was designed to evaluate whether GMAP achievement levels could be considered comparable to the achievement levels associated with Georgia Milestones. This could provide evidence in support of achievement-level comparability—a requirement for innovative assessments under the IADA.



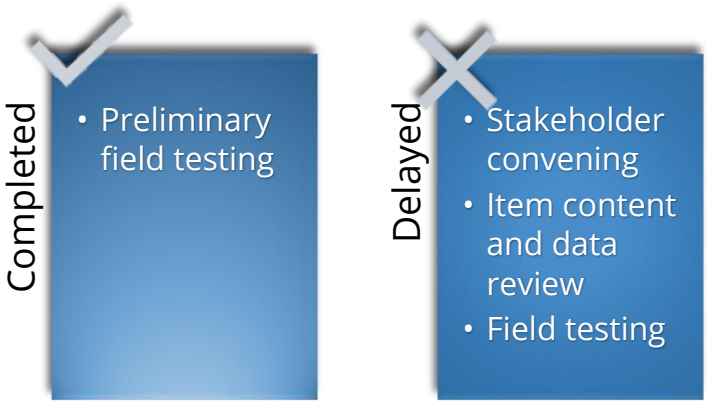
In addition to delaying their schedule, GMAP was forced to redesign already-scheduled meetings to accommodate the virtual formats that became the norm in 2020. Most meetings were able to proceed in this format. NWEA continued with efforts to develop family reports and test item development, and with support from the Walton Family Foundation, GMAP gathered feedback on family reports through focus groups in January 2020. This work would continue through the 2019–2020 school year.

During the 2020–2021 school year, NWEA plans to provide professional learning support to schools, continue work on the GMAP family report, and continue content development in ELA, Math, and Science. At the June 2020 TAC meeting, GMAP and NWEA noted that their plans for the field test design and adaptive test design were still underway. Additional research studies in 2021 are planned to inform these designs. Key to these studies will be establishing sufficiently large samples; stable item parameter estimation for their operational item bank will require many student responses to each item.

Putnam County Consortium

At the December 2019 TAC meeting, the Putnam County Consortium and Navvy shared that the schools participating in the consortium have been utilizing the Navvy assessment system since the 2017–2018 school year. Each year, additional development has added to the pool of Navvy test items and has supported expansion of the tests to additional grades. The

assessment system is currently available to students in grades 3–8 and high school in ELA and Math.



Putnam had hoped to establish comparability between Navvy and Georgia Milestones in Summer 2020, with science content expected to come online in time for field testing in 2021–2022. Although students were able to take the Navvy assessment in ELA and Math during the 2019-2020 school year, the data are incomplete, since the COVID-19 pandemic shifted instruction to at-home virtual classrooms. At the June 2020 meeting, the TAC suggested that these data could be used to make some comparability predeterminations, so that any necessary changes could be implemented for the following school year. The TAC also encouraged Putnam to establish comparability via achievement level descriptors.

For the 2020–2021 school year, the Putnam County Consortium plans to further scale up the Navvy Assessment System. They will attempt to recruit additional districts to participate in the pilot; Putnam also plans to engage with stakeholder groups including strong representation from historically marginalized populations. Stakeholders will ultimately provide feedback on the assessment system as well as participate in item content and data review meetings. Finally, the Putnam County Consortium will refine its plans for collecting validity evidence—including analyses of test content, response processes, test consequences, and relationships with other variables such as expert diagnoses.

Technical Assistance Services Provided

WestEd provided technical assistance to each consortium during the contract period. Putnam and GMAP were each allotted 114 technical assistance hours to be used at their discretion (GMAP used 30.5 hours; Putnam used 71.5). WestEd worked with the consortia on annual reporting, administration best practices, operationalizing an assessment system, implementing testing policies, engagement of stakeholder groups, and processes for holding content and bias reviews. WestEd also served as a liaison between the consortia and GaDOE when questions about Georgia Milestones policies and documentation arose. Additionally, WestEd planned and hosted two TAC meetings in this inaugural year of the Georgia IAPP. Each consortium met with the TAC for one day at each meeting. Participant districts, their

test development partners, WestEd, GaDOE, and the Governor's Office of Student Achievement (GOSA) took part in the TAC meetings. The first meeting was convened December 9–10, 2019, in Atlanta, Georgia, and the second meeting was postponed briefly and ultimately hosted as a virtual meeting from June 29–30, 2020.

Taking Stock of COVID-19's Impact on the Georgia IAPP, 2020-2021

In spring 2020, schooling for many students across the U.S. simply ground to a halt. Summative tests were cancelled and so were their through-year counterparts such as Navvy and GMAP. Both consortia took the opportunity to engage in further research to support their assessments, content and bias reviews, and other activities that did not require students to be in schools. Neither consortium was concerned that the end of the initial 5-year IADA period could arrive before their tests were used for summative purposes. That outcome seems somewhat more likely now. **Because neither group will finish field testing before spring 2022, neither will be able to administer an operational assessment for comparability purposes throughout a school year until 2022-2023. At that point it will be extremely difficult to secure approval for the 2023-2024 school year since comparability analyses could not be conducted until spring testing was complete in June. That leaves 2-3 summer months, at best, (1) for the consortia to present comparability results to the TAC, incorporate feedback, and submit results to GaDOE, and then (2) for GaDOE to convene the necessary expert panels, review evidence (getting clarification from the consortia as needed), make approval determinations, and change each participant district's contract. WestEd is working closely with GaDOE to devise a practicable solution to this issue, but GaDOE cannot rush through a review given the comprehensive nature of the assurances (see p. 40 in the Appendices).**

Finally, COVID-19 also promises to significantly disrupt at least one more testing cycle – spring 2021. This is unfortunate timing for the pilot participants, as both were planning to field test items this year. Both continue to move forward with tentative plans, but there are major concerns about the quality of data generated in spring 2021. There are also no hard and fast rules clarifying the amount of student absence that is tolerable, such that item calibration and other core psychometric analyses can go forward.

WestEd's recommendation is to lean on the like-minded fields and organizations that have been developing methods for years to handle what will ultimately manifest as large-scale attrition. The preferred approach from our perspective would be to treat spring 2021 as an extensive, systematic, missing data problem. Then, we recommend following guidance from the U.S. Department of Education's Institute of Education Sciences (IES), which recently developed heuristics for handling missing data in rigorous experimental or quasi-

experimental studies. In some cases, the influence of missing data can be minimized through weighting or multiple imputation, such that unbiased parameter estimates can be drawn from datasets with high missing rates.

More specifically, the consortia could review the most recent version of the What Works Clearinghouse Group Design Standards (missing data is discussed on p. 33), which essentially represents IES's current thinking on advanced topics like imputation. Conveniently, the Group Design Standards offer simple metrics and cutoffs (e.g., missingness above 15% cannot be ignored), which could be applied to spring 2021 data. GaDOE and the TAC may decide that, with some safeguards, following the missing data methods that are required of large-scale randomized trials will suffice for the Georgia IAPP.

Lessons Learned and Next Steps

The three lessons below all qualify as major recurring themes over the course of 2019–2020. None are particularly technical in nature; psychometric concerns that surfaced and were satisfied may have been lessons learned, but our intent in calling out the points below is to highlight particularly thorny issues that touch those who build innovative tests, those who put them to use, those who act on the data tests generate, and those responsible for monitoring the system's health over time.

In keeping with the knowledge-sharing principles that animate this demonstration authority, the issues below reflect concerns that are common across multiple IADA-approved states. Any examples we cite are Georgia-specific, but the themes they represent are, in our experience, ubiquitous. Finally, whenever possible, we couple the issues with potential remedies or new paths to consider. Not every concern has a solution. That said, none of the problems seem intractable. Quite the opposite; in a year when initiatives and industries stalled completely, implementation of the Georgia IAPP is progressing steadily.

Lesson 1: The Resources Required

- It is widely acknowledged that developing and scaling a truly innovative assessment system is not a break-even proposition. Intuitively, organizations must spend new dollars to create new programs.
- Less-widely understood are the varied assets vendors and state department staff will need to bring to this work. Vendors of course need an abiding commitment to experimentation, but they must also develop the capacity to think like a traditional assessment program. New assessment models under IADA must strike a delicate balance: breaking new ground to solve high-leverage problems of assessment practice without compromising fairness and transparency. Innovative assessments are still high-stakes assessments, governed by the bedrock standards that have supported educational measurement for decades. This means that security, accessibility, and appropriate accommodations are as important within IADA as they are without it.
- Through our work with four IADA-approved states, we have learned that state education departments will deal with an entirely different challenge: they already have an assessment program to run. So, asking state department staff to also shepherd along a new assessment program (intended to supplant theirs) without sufficient discretion, preparation, or flexibility could put both programs at risk. To protect the integrity and validity of testing programs and test scores, we recommend keeping open the lines of communication between policymakers and state education agencies.

Lesson 2: The Major Hurdles

- Perhaps because the educational measurement field has been particularly vocal on IADA, over the past year we have observed a persistent, often disproportionate level of concern attached to once-esoteric topics such as adaptive algorithms and score comparability. Under almost any other circumstance, serious attention to these issues would be welcome. Under IADA, there will be bigger fish to fry.
- Consider, for example, that even large technology companies have been flummoxed by the requirements of online testing in the best of times, in relatively controlled environments (schools) with known technology capacity. If, in the COVID era, remote proctoring becomes the norm rather than the rare exception—even briefly—how can the nimble, innovative, but comparatively underresourced assessment startup ensure error-free administrations? With summative testing facing more than the customary amount of public skepticism, the answer to this question should be important to innovative and traditional programs alike. By comparison, performance-level comparability sounds positively sortable.
- Similarly, a fixation on psychometric comparability can distract test developers from the many other minimum requirements of a summative statewide testing program. It is well and good for ED to keep the standards for IADA entry as lenient as possible, but statewide testing programs are still subject to state law. Pilot participants should know that psychometric comparability is not the only criterion a state must consider when authorizing an accountability test. In Georgia, as in many other states, there is more to it than that (See Appendix 2 for the assurances associated with Georgia's IAPP).

Lesson 3: The Upside and the Downside of Competition

- The Georgia IAPP is innovative not only in the through-year assessment systems it will produce, but also in the intrastate competition it has promoted through IADA. Georgia is a test case for this model, which has not been adopted by any other IADA-approved state. The rationale is straightforward; a competitive field will raise the level of play, so to speak, and ultimately more students in more schools will be assessed with better instruments. However, competition in this setting introduces some discontents.
- **First, it becomes necessary for the state department of education to adopt a rigorously neutral stance toward all participants and act with an overabundance of caution for at least the five years it will take the consortia to mature and scale. Third-party TA providers will have to step in to fill the void, and while we have been thrilled to support the state of Georgia, we would be doing a disservice to our client if we did not point out that the current arrangement could put GaDOE in a difficult position.**
- Second, a bona fide competition necessarily undermines one of the central goals of IADA: the sharing of knowledge between diverse organizations pursuing the same goal (advancing student learning) in vastly different ways. Strict—and appropriate—confidentiality protocols limit what Putnam can learn from GMAP, and vice-versa. We would recommend taking some affirmative steps toward opportunities for collaboration (e.g., a jointly-hosted ideas summit). Each consortium, not to mention the IAPP itself, stands to benefit.
- Lastly, when the end of the IADA period brings this competition to a close, Georgia will need to accept one assessment system. That means some districts will spend at least five years implementing and advocating for the system they have chosen and then will be forced to adopt the one they have not. So, while a quick statewide embrace of this competition's "winner" is not impossible to imagine, it is also not very easy to imagine. Backlash is the last thing IADA's architects want; to avoid it, districts from the "losing" side in Georgia will need assurances that the new statewide system will advance their interests.

APPENDICES

Appendix 1A

GEORGIA INNOVATIVE ASSESSMENT PILOT PROGRAM

December 2019 Technical Advisory Committee (TAC) Meeting Report

Putnam County Consortium

January 25, 2020
Submitted by:
WestEd
730 Harrison Street
San Francisco, CA 94107

DECEMBER 2019 TAC MEETING REPORT FOR THE PUTNAM COUNTY CONSORTIUM

INTRODUCTION

The Georgia Innovative Assessment Pilot Program (IAPP) Technical Advisory Committee (TAC) meeting was convened on December 9, 2019, in Atlanta, Georgia. Attendees included members of the TAC; the Putnam County Consortium (Putnam Consortium); Navvy Education, LLC; the Georgia Department of Education; and WestEd. This report provides an overview of the topics discussed and a description of the resulting key takeaways and action items from the meeting.

INTRODUCTION TO NAVVY ASSESSMENT SYSTEM

Description

The Putnam Consortium and Navvy Education provided an overview of the purpose, design, and implementation of the Navvy Assessment System (Navvy). The purpose of this topic was to provide the TAC with introductory information about the assessment system. Navvy Education assessment representatives also shared examples of the user interface with the TAC.

TAC Discussion and Recommendations

The TAC took this opportunity to learn more about Navvy by asking questions about assessment design and implementation. The TAC had recommendations in three areas, based on this discussion.

First, the TAC posed questions about the impact that this assessment system could have on teaching. The Putnam Consortium shared that it has received favorable feedback about the utility of formative information that Navvy provides. The TAC noted that, once accountability is introduced into the system, the Putnam Consortium may want to conduct additional research into how the standards are taught to students, to ensure that the standards are not presented more prescriptively once high stakes are attached to Navvy.

Second, the TAC complimented Navvy's user interface, noting that it is a mechanism to encourage teachers, parents, and administrators to review, understand, and use the Georgia content standards. Discussion of the user interface included discussion of parent access to the system. Currently, parents can use their student's username and password to access the student's records. The TAC suggested that, in the future, a rostering formula-based username and password for parents to access the system would be beneficial.

Finally, the TAC learned more about the types of items administered via Navvy. Currently, there are multiple-choice and multiple-select item types. The TAC suggested that, in the future, additional item types should be considered, because Georgia Milestones also administers technology-enhanced items.

Comparability Plans for the Putnam Consortium

Description

The Putnam Consortium discussed approaches to creating annual summative determinations as well as to establishing comparability with Georgia Milestones.

TAC Discussion and Recommendations

The TAC emphasized that the Putnam Consortium should focus on how to establish comparability in achievement-level classifications in order to move forward to implementation under the rules of the IAPP. For example, Navvy could create four achievement-level classifications and use linear or logistic regression methods to maximize classification consistency relative to Georgia Milestones achievement levels. The TAC recommended that the consortium align its performance classifications with existing Performance Level Descriptors for Georgia Milestones, to the extent practicable. The TAC also suggested that the consortium consider weighting its measures to align with the current Georgia Milestones blueprint, although this would not be required in order to establish comparability.

Students taking the Navvy assessment are given three attempts to show that they have reached proficiency relative to a given standard. The TAC discussed the number of attempts that should be used when calculating comparability to the Georgia Milestones assessment, and recommended that the consortium calculate scores for comparability analyses at the second attempt. The TAC agreed that utilizing results from a sample of 300–400 students per grade and content area would be sufficient to establish comparability, assuming that the distribution of Navvy examinees is similar to that of Georgia Milestones examinees. For future meetings, the TAC is interested in seeing more information on pacing and sequencing — that is, when attempts for each standard are administered across grades, subjects, and schools.

Writing Assessment

Description

The Putnam Consortium provided an update on development and implementation of writing assessments within the Navvy system.

TAC Discussion and Recommendations

Navvy currently assesses writing through extended-response items, whereas Georgia Milestones also administers multiple-choice writing items. The TAC advised that the writing standards addressed must be tested at the same depth and breadth (within grade bands) as in Georgia Milestones. The TAC suggested that the writing assessment be included in students' ELA scores and utilized when establishing comparability.

Implementation Supports for Member Districts

Description

The Putnam Consortium provided an overview of its plan to provide supports to districts implementing the Navvy Assessment System.

TAC Discussion and Recommendations

The TAC noted that existing communities of practice have provided useful resources to diverse consumers of educational assessments. For example, the Smarter Balanced consortium provides a "digital library," which functions as a repository of assessment resources that have been vetted by experts in the educational assessment field. Additionally, the Advanced Placement assessment program has a teacher-led community of practice, in which members share lessons and other tasks that they have successfully used in their classrooms. These communities can serve as models for the Putnam Consortium to reference. The TAC noted that any resources being provided to districts should be vetted by experts in educational assessment.

Next Steps

Spring/Summer 2020 TAC Meeting

The next TAC meeting will focus on a concrete, near-term task: IADA Annual Performance Reporting. IAPP participants' reports are due to the Georgia Department of Education in summer 2020, so the next TAC meeting will generate feedback for the Putnam Consortium, to inform the Annual Performance Report (the report template is included as an attachment to this report). In particular, we hope to focus on the infrastructure and project management required to successfully deliver a large-scale summative testing program (e.g., quality assurance, test security, accommodations, scoring and reporting).

Future Work

The TAC suggested that long-term planning and analysis should include the following items:

- Provide descriptive data giving information on the number of attempts per student per standard, along with mastery rates
- Provide information on when districts are administering Navvy and Georgia Milestones (within the testing window), to gauge whether test timing could impact comparability
- Provide demographic data across all participating districts, with comparison to the demographics of the state of Georgia

Appendix 1B

GEORGIA INNOVATIVE ASSESSMENT PILOT PROGRAM

DECEMBER 2019 TECHNICAL ADVISORY COMMITTEE (TAC) MEETING REPORT

Georgia MAP Assessment Partnership

January 25, 2020

Submitted by:
WestEd
730 Harrison Street
San Francisco, CA 94107

DECEMBER 2019 TAC MEETING REPORT FOR THE GEORGIA MAP ASSESSMENT PARTNERSHIP

INTRODUCTION

The Georgia Innovative Assessment Pilot Program (IAPP) Technical Advisory Committee (TAC) meeting was convened on December 10, 2019, in Atlanta, Georgia. Attendees included members of the TAC, the Georgia MAP Assessment Partnership (GMAP Partnership), Northwest Education Association (NWEA), the Georgia Department of Education, and WestEd. This report provides an overview of the topics discussed and a description of the resulting key takeaways and action items from the meeting.

OVERVIEW OF THE GMAP THROUGH-YEAR SOLUTION

Description

The GMAP Partnership and NWEA presented an overview of the GMAP through-year model. The NWEA presentation provided an overview of the model as well as the timeline for development. NWEA explained how its through-year model compares to traditional summative tests, as well as to its MAP Growth assessment. Details on the design of the through-year model were presented, providing the TAC with information on the computer-adaptive testing algorithm used to route students to items.

TAC Discussion and Recommendations

During its presentation, NWEA explained that the adaptive algorithm accommodates students testing off-grade, providing students with items that relate to the on-grade content standards. The TAC recommended that the GMAP Partnership gather evidence showing how off-grade-level items are aligned to on-grade-level content. The TAC also suggested using the adaptive engine to select performance tasks, particularly in the math domain.

For reading assessments, the TAC discussed how the adaptive engine would function for off-grade, passage-based items. Ideas included developing multiple versions of each passage, with differing complexities; developing differing prompts for the same passage; and developing off-grade items for a particular passage to be field tested. The TAC noted that student ability estimates (i.e., thetas) should not be too dependent on a single reading passage.

The GMAP Partnership asked TAC members to reflect on how the current through-year test design addresses the intent of the Every Student Succeeds Act. The TAC advised that the through-year design should focus on both the breadth and the depth of the state content standards. The TAC also noted that if the test blueprint remains the same across administrations within a school year, creating the required summative score that needs to be reported may be easier. However, maintaining identical blueprints across the year may not be required, and allowing the blueprint to shift across administrations may provide more actionable information.

Lastly, the TAC recommended that communication to teachers address how to use the data produced from the various testing events throughout the year. For example, because 60 percent of all items administered throughout the year must be on grade level, the third testing event for students with below-grade proficiency may contain mostly items that are on grade level (assuming that prior testing occasions contained larger shares of below-grade-level items). Teachers should have guidance on how to interpret and use the data from these comparatively difficult tests.

Comparability to Georgia Milestones

Description

NWEA described a planned research study that will gauge the value of achievement level descriptors (ALDs) for providing feedback to teachers and students. The use of ALDs to establish comparability to Georgia Milestones was also discussed.

TAC Discussion and Recommendations

The TAC recommended that the GMAP Partnership utilize ALDs for establishing comparability; however, the research agenda is not required in order to establish comparability under the Georgia IAPP. In order to establish comparability, the GMAP Partnership should demonstrate that students' achievement-level classifications are comparable to Georgia Milestones. Evidence of comparability at the raw score or scale score level will not be necessary.

The TAC also noted that, to establish comparability, the GMAP Partnership will also need to produce a literacy measure and a growth indicator. It is important to emphasize, however, that the GMAP Partnership does not need to establish comparability between its growth metric and the state's growth metric (student growth percentiles). Rather, the GMAP Partnership should adopt or develop a growth model that aligns well with NWEA's through-year assessment. The TAC also noted that the GMAP Partnership's literacy measure should be related to Georgia's literacy measure (Lexiles), but evidence of achievement-level comparability will suffice for the IAPP.

Incorporating the RIT Scale

Description

The GMAP Partnership described for the TAC how it plans to include RIT scores (generated for MAP Growth assessments) in its through-year assessment model, in order to provide Georgia students with norm-referenced information.

TAC Discussion and Recommendations

The TAC noted that there are compelling reasons for incorporating the RIT scale into NWEA's through-year assessment model. MAP Growth scores will provide a familiar anchor for students taking a new summative assessment in lieu of Georgia Milestones. However, the GMAP Partnership's priority should be the development of a new through-year assessment, not the provision of RIT scores. Therefore, field-test designs and calibration and equating procedures should not compromise the through-year assessment scale in order to

accommodate the RIT scale. For example, if the through-year assessment includes performance tasks and MAP Growth does not, putting through-year assessment on the RIT scale may not be advisable.

Scaling to Statewide Implementation

Description

This discusssion focused on how the GMAP Partnership — a consortium of districts in Georgia — would ultimately be able to transition to a full statewide assessment program.

TAC Discussion and Recommendations

The TAC suggested that the GMAP Partnership develop readiness criteria for districts, articulating the key features that successful districts exhibit. Additionally, the TAC recommended researching lessons learned from the Race to the Top large-scale assessment consortia (Smarter Balanced and PARCC). The TAC noted that when multiple parties attempt to reach an agreement, it is difficult for all preferences to be accommodated. As any assessment system becomes more customized to meet varying preferences, there are implications for cost, development time, and assessment quality and validity.

Next Steps

Spring/Summer 2020 TAC Meeting

The next TAC meeting will focus on a concrete, near-term task: IADA Annual Performance Reporting. IAPP participants' reports are due to the Georgia Department of Education in summer 2020, so the next TAC meeting will generate feedback for the GMAP Partnership, to inform the Annual Performance Report (the report template is included as an attachment to this report). In particular, we hope to focus on the infrastructure and project management required to successfully deliver a large-scale summative testing program (e.g., quality assurance, test security, accommodations, scoring and reporting).

Future Work

The TAC suggested that long-term planning and analysis should include the following items:

- Provide documentation showing the alignment between the through-year assessment's ALDs and the Georgia ALDs
- Provide documentation showing the alignment of the through-year assessment's DOK levels to Georgia Milestones
- Provide a high-level description of the field-test plan
- Provide Georgia Milestones score comparisons across participating districts, with demographic data included
- Provide sample reports for very high-performing and very low-performing students, to show how interpretable data can be generated from different sets of items delivered

Appendix 1C

GEORGIA INNOVATIVE ASSESSMENT PILOT PROGRAM

JUNE 2020 TECHNICAL ADVISORY COMMITTEE (TAC) MEETING REPORT

Putnam County Consortium

July 14, 2020

Submitted by:
WestEd
730 Harrison Street
San Francisco, CA

JUNE 2020 TAC MEETING REPORT FOR THE PUTNAM COUNTY CONSORTIUM

The Georgia Innovative Assessment Pilot Program (IAPP) Technical Advisory Committee (TAC) meeting was convened on June 30, 2020. The meeting was held virtually via Zoom video conferencing. Attendees included members of the TAC, the Putnam County Consortium (Putnam Consortium), Navvy Education, LLC, the Georgia Department of Education (GaDOE), and WestEd. This report provides an overview of the topics discussed and a description of the resulting key takeaways and action items from the meeting.

UPDATE ON PUTNAM CONSORTIUM AND NAVVY ASSESSMENT SYSTEM

Description

The Putnam Consortium and Navvy Education provided an overview of the Navvy assessment system and a progress update on their timeline, which has been impacted by the COVID-19 pandemic. The TAC also engaged in discussion around determining proficiency levels and methods for establishing comparability using achievement level descriptors (ALDs).

TAC Discussion and Recommendations

Navvy explained their relationship with Putnam County and their joint desire to test students in a standard-by-standard fashion and to measure overall proficiency with reference to students' mastery of each standard. Member districts joined because they have a shared interest in assessing students in this way. Navvy was built as a formative assessment tool for teachers to use throughout the school year in their classrooms to support teaching and learning, with the eventual goal of using Navvy in lieu of Georgia Milestones (hereafter "Milestones") for state accountability purposes.

The TAC discussed the decision consistency and decision accuracy of proficiency determinations under Navvy vis-à-vis Milestones. The TAC noted that Navvy is able to use the state achievement levels for comparability. They recommended using the Milestones technical report to get the data for a baseline for comparison. Because Navvy is so targeted to standards, comparisons to Milestones may be most productive at a higher level, such as the ALDs. The TAC suggested exploring achievement-level alignment between Navvy and Milestones by looking at what Georgia's ALDs say students should be able to do at each grade level and in each subject. The TAC supported Putnam's idea to use a cluster-analytic approach to determine if there are patterns that characterize where students are landing (in terms of their Navvy scores). One approach would be to pre-assign cluster centroids based on Milestones achievement levels. Navvy might also create its own achievement levels and then align them to the state's ALDs to demonstrate comparability. Another suggestion was to validate the standard-by-standard assessment approach Navvy is adopting via a Rasch or other Item Response Theory (IRT) model, using each student's most recent valid assessment score.

The TAC addressed the challenge of fitting the results of the Navvy assessment into Milestones ALDs, because those ALDs address what students can do globally and are typically determined based on summative tests. One suggestion is to create policy labels and descriptors and link them to Milestones' labels and policy descriptors. Alternatively, Navvy could explore expressing ALDs as the probability that a student has mastered a skill.

The TAC discussed if this assessment model measures long-term mastery of a standard, because students aren't retested on a standard once they show proficiency. A possible future research study could be to retest students quite a few months after a proficiency score is received to determine whether that score is still accurate.

The Putnam consortium discussed the impact of the COVID-19 pandemic on testing in Spring 2020. Students in Georgia did not take Georgia Milestones, nor were they able to complete testing on the Navvy assessment, since they were not in classrooms. This will put the Putnam Consortium behind in their anticipated timeline but still allows enough time for comparability to be established within the pilot program's window. The TAC suggested that a "pre-comparability" study could use the data Navvy was able to collect in the 2019/20 school year and the Milestones results from the 2018/19 school year.

Strategies for Scaling

Description

The Putnam Consortium and Navvy Education shared their current plans for scaling the assessment system. The TAC provided feedback on how to engage additional districts and stakeholders in the consortium's activities in order to help grow membership and increase participation.

TAC Discussion and Recommendations

The TAC provided recommendations to the Putnam Consortium on how to scale up the assessment system throughout the pilot program. The TAC suggested offering the practice tests publicly so prospective districts can experience the test for themselves. If they have a positive experience with the practice test, they may be more inclined to want to join the consortium. One challenge to recruitment is the state's new interim assessment tool that is offered free of charge to districts, called Beacon. Putnam has received feedback from stakeholders that they think Navvy and Beacon are similar tools that will provide similar results. The TAC suggested that district leadership in the Putnam Consortium help communicate to the public how the assessments differ.

The Putnam Consortium plans to increase communication with districts and stakeholder groups as a part of their scaling activities in the coming year. The TAC encouraged Putnam to engage the members of their current committees and groups as advocates for the Navvy Assessment. They also suggested engaging stakeholders from various organizations in the state as a part of their various committees and feedback groups. The TAC emphasized that these stakeholder groups should include representation from historically marginalized populations.

The TAC also made suggestions on methods for marketing materials. They suggested that materials emphasize teacher utility and include information on how the assessment provides information on each specific academic standard. They provided a list of key words and phrases that should be incorporated into Putnam's materials, including: "fully-aligned," "actionable," "instructionally relevant," "just in time," "immediate," and "student-focused." Specifically, the TAC discussed developing a brochure with a table that compares Navvy to other assessment tools.

Strategies for Continuous Improvement

Description

The Putnam Consortium and Navvy Education further discussed a plan to engage with stakeholders and key experts in order to receive meaningful feedback on the Navvy Assessment system. They asked for feedback from the TAC on their plan.

TAC Discussion and Recommendations

The TAC suggested that the panels and groups that Putnam engages should have a clear definition and purpose. When meeting with the panels, Putnam should make sure to level-set with the participants so they know what type of feedback is being solicited. Participants will need an understanding of what kind of change they can actually make, and that there are some restrictions based on test design or by state and federal law and policy. The TAC also suggested to pare down the number of groups in the current plan, since participants in each group overlap. They encouraged Putnam to ensure the makeup of each panel is diverse — making sure to engage minority groups, including people with disabilities. They should also be sure to be very transparent about the changes that are made as a result of the panel feedback.

The TAC also suggested soliciting comment from various stakeholder groups when issues arise. To do this, the Putnam Consortium would gather a list of organizations and groups to engage with on an as-needed basis. When issue arise, they would reach out to all the groups on the list, asking for their feedback.

The TAC discussed how to elicit feedback from parents and students. They suggested that they could have parent representation on the leadership panel. Information on their experiences could also be provided through teachers. This would be particularly useful for collecting information on student experiences, because it is not desirable to have young students sit on an advisory panel of this nature. They may want to consider having high school student representation on a policy panel. Another way to involve students is to conduct focus groups about the future of assessment.

Strategies for Collecting Validity Evidence

Description

The Putnam Consortium and Navvy Education presented their planned activities to help build validity evidence. Studies were presented for five areas — Evidence Based on Test Design, Evidence Based on Response Processes, Evidence Based on Internal Structure, Evidence based on Relationships to Other Variables, and Evidence Based on Test

Consequences. The TAC provided feedback in a few of these categories, reminding Putnam to only do what is required of the innovative assessment pilot program at this point, so as not to overcommit themselves to too many studies. At minimum, they must show comparability with Milestones.

TAC Discussion and Recommendations

Evidence based on Test Content

The TAC discussed that the evidence needed to show validity based on test content should include an alignment study. GaDOE confirmed that, as a part of the Innovative Assessment Pilot Program, the state will fund an external alignment study in a future year.

To make claims on alignment until an alignment study is conducted, there needs to be an external, independent validation that content is aligned to the standards. The current process of how items are written and reviewed internally should be documented. To further the evidence on validity, the Consortium should engage teachers as independent reviewers from the participating districts to review the items and confirm their alignment. The TAC also suggested providing more information on the consistency of the content representation to which students are exposed. The Putnam Consortium will bring this as a topic to revisit in the December 2020 TAC meeting.

Evidence Based on Response Process

The Putnam Consortium presented an outline for conducting cognitive labs that would collect evidence on response process. The TAC affirmed that this evidence is meaningful because it is needed to support the assessment's claim that students are engaging in a certain cognitive process. To ensure findings are generalizable, data should be collected across sub-groups. Log data should also be mined to collect information, such as speed of response. The TAC also suggested that an informal cognitive lab could be conducted by asking teachers to pilot items and then ask students questions about the items they took. The TAC suggested reviewing existing literature on response process (suggested authors are Zumbo & Hubley, Leighton, and Pellegrino).

The TAC shared that Putnam's current sampling plan for the cognitive lab plan is slim and would likely not produce enough evidence. Increasing the number of standards addressed and conducting the study over multiple years, with new standards each year, would yield stronger evidence. A strategy should be defined for sampling the standards, such as identifying foundational standards that vary across depth of knowledge (DOK) levels.

The TAC also suggested that response data be reviewed against the type of device students are taking the assessment on. Navvy does not recommend using a mobile device, but it is not prohibited. By looking at this data, Putnam can consider if there are any response processes that introduce additional errors responding due to the device the assessment is administered on.

Evidence Based on Relationship to Other Variables

Consistency with External Expert and Model-based Diagnoses. The TAC shared that evidence produced from this study may not be strong because it is a small, selective sample.

They also noted that this will be just one of many pieces of evidence that they plan to use to produce evidence of validity, and that some pieces will inevitably be stronger than others. The TAC suggested that Putnam may want to explore other methods of triangulating this data. One suggestion is to administer test questions in an open-ended format to students instead of multiple choice in order to see if they would produce the same answers. Another option would be to develop behaviorally anchored rating scales (BARS), which teachers could fill out for students who also took the Navvy assessments. Those BARS would function similarly to course grades in a concurrent validity analysis (that is, we would expect them to correlate positively with Navvy scores). While a BARS would not take long for a teacher to fill out (roughly 30 minutes per student), it would take some time to develop, since the behaviors would have to be specific, explicit, differentiated, and exhaustive enough to capture the same information Navvy performance levels capture.

Consistency with Other Measures. The TAC suggested that they build into this study something in which students are given Navvy twice, before and after instruction, to see how their score changes. There need be no limitations on what the variables are. They may want to use a multi-method multi-factor model to look at correlations between classroom grades and assessment results (if teachers are not using Navvy in their students' grades). Showing that there is a strong correlation between the two will also help with marketing the assessment to other districts. The TAC noted that this would not be a good source of psychometric evidence, however, because grades and assessment results are not measuring the same thing.

The TAC discussed using reliability evidence for validity by using domain reliability within trait reliability and factor analysis. This captures reliability and validity at the same time and would contribute to the validity evidence. The goal of this study would be to end up with reasonable correlations with Milestones. They won't be perfect correlations, because the test allows for interventions that break the cycle of the traditional summative results.

Putnam asked the TAC if validity evidence based on the relationship between Milestones and Navvy should show correlation with the total score or by domain. The TAC recommends running the correlations at the total score level. They still encourage exploring the correlations at the domain subscores and competency rates to see what they find. It would be valuable to set up a theory ahead of time that explains what Putnam expects to see.

Evidence Based on Test Consequences

The TAC reminded Putnam that, when communicating information about the validity of the assessment, they should be sure to have a statement of the intended score meaning — what scores are supposed to mean. The TAC recommended scaling back the number of intended consequences/effects of the assessment system on which to gather information. The TAC recommended focusing on finding out if and how teachers are using Navvy to improve their instruction. Measures should be matched with students and not across districts. The TAC recommended putting more emphasis on subgroups and putting more effort into looking at subgroup analyses, differential impact, and differential access.

The TAC noted that a common issue with state summative assessments is the need for results to be provided quickly, and therefore content needs to be lower-level. The TAC would like to know if Navvy will have this same problem or if, instead, students are given the opportunity to demonstrate level three and level four thinking. In the future, the TAC would also like to see some of the test forms and the results from those forms. They also suggested providing longitudinal data to help support claims that the assessment is contributing to improvement of learning.

Pandemic Impact on Pilot Timeline and Activities

Description

The Putnam Consortium and Navvy Education discussed the impact that the COVID-19 pandemic has had on their timeline to conduct pilot activities. They still plan to conduct a data review with the data that was collected before schools ceased in-person instruction. They also discussed with the TAC what policies may also need to be rethought due to the unknown impact the pandemic will have on classroom structures.

TAC Discussion and Recommendations

The TAC affirmed that the Putnam Consortium should still move forward with data review and data calibrations with the data that was collected this year. When more complete data is collected, they should run another analysis.

The TAC encouraged the Putnam Consortium to rethink their test security policies for students who will be attending school from their homes in the next school year. Test security will need to be reconsidered, especially since the data that teachers receive from this assessment could be very helpful for them during a tumultuous time. The TAC encouraged Putnam to think about which elements of their model are preferable and which are negotiable in order to rethink plans for the coming school year. Some suggestions on how to have students test from home included having a way for their browsers to be locked down during test taking, as well as having students sign an affidavit acknowledging that they understand test security rules.

Other ideas for the Putnam Consortium to consider were around limiting exposure to the item bank. One suggestion was to allow only a single instance of the assessment for each standard or providing a window of time (e.g., 30 days, 45 days) before students can retake an assessment. Another suggestion was to hold items tested this year for one to two years before putting them back into the form pull rotation (it is unknown how peer review would view this approach). Besides peer review, another drawback to at-home testing is issues around equitable access to computers and internet.

Data Review Procedures

Description

The Putnam Consortium and Navvy Education discussed their planned procedures for data review. The TAC provided guidance on quality control, screening data for non-effortful responding, and planning for data review panels.

TAC Discussion and Recommendations

The TAC recommended that the Putnam Consortium and Navvy ensure that they have the resources planned to ensure that quality control measures are in place. They also recommended they have protocols set up specifically for protecting personally identifiable information that data reviewers could possibly access.

The TAC recommended that when flagged items are reviewed, particular attention should be paid to the items in which the percentage of students choosing the correct response falls below change. Sometimes it is a very good item, but it still gets flagged because it is a more difficult item, but not a problematic one.

Navvy plans to screen response data for non-effortful responding. The TAC suggested that they look into literature about that topic to weigh various options for establishing reasonable cut-offs for response time. In particular, Steve Weiss has written about a few different criteria Navvy can consider. There are some districts that are not currently administering an assessment for every standard to their students. The TAC recommends running data analysis on the entire system and on the subset of schools that are administering an assessment for every standard. This way comparisons can be made between the two.

When planning for data review, the TAC reminded the Putnam Consortium that they should ensure their review panels are diverse and are representative of minority populations. Additionally, they advised that ground rules be provided to participants and ensuring they understand that they are tasked with making dichotomous decisions. When there is a large item bank, review meetings can take a long time, and this would dissuade participants from trying to rewrite items. They should also keep records of the decisions the panels make on every item, so it can be referenced in the future, if needed.

NEXT STEPS

Future Work

At the conclusion of the meeting, the TAC requested the following information during future TAC meetings:

- Utilization and implementation data (How many students are taking multiple attempts on an assessment? How is the assessment being used across districts, grades, and subject areas?)
- Examples of how results will be communicated and presented to educators, parents, and students
- Plans for gathering feedback from stakeholder groups, particularly teachers and parents

During the TAC debrief with GaDOE and WestEd, the TAC also recommended that each consortium discuss the following topics in future TAC meetings:

- Comparability within the assessment system
- Updates on any independent alignment studies that have been conducted
- Plans for score reporting

Appendix 1D

GEORGIA INNOVATIVE ASSESSMENT PILOT PROGRAM

JUNE 2020 TECHNICAL ADVISORY COMMITTEE (TAC) MEETING REPORT

Georgia MAP Assessment Partnership

July 14, 2020

Submitted by:
WestEd
730 Harrison Street
San Francisco, CA 94107

JUNE 2020 TAC MEETING REPORT FOR THE GEORGIA MAP ASSESSMENT PARTNERSHIP

INTRODUCTION

The Georgia Innovative Assessment Pilot Program (IAPP) Technical Advisory Committee (TAC) meeting was convened on June 29, 2020. The meeting was held virtually, via Zoom video conferencing. Attendees included members of the TAC, the Georgia MAP Assessment Partnership (GMAP Partnership), Northwest Education Association (NWEA), the Georgia Department of Education (GaDOE), and WestEd. This report provides an overview of the topics discussed and a description of the key takeaways and action items resulting from the meeting.

UPDATE ON CONSORTIUM ASSESSMENT SYSTEM

Description

The GMAP Partnership and NWEA presented updates on their work on the GMAP through-year assessment. The partnership provided information about consortium membership, assessment development activities that have been completed, and plans for future activities. The TAC was asked to provide feedback on the decision-making process for the field-test plan and on the process GMAP is following to select among candidate adaptive test designs.

TAC Discussion and Recommendations

NWEA first summarized progress on test development over the past year. In that time, NWEA project staff have focused on planning, item development, and item reviews. Since the previous TAC meeting in December, they have directed additional attention to the design of individual student reports. In collaboration with the Walton Family Foundation, focus groups were conducted to gather input on student reports. This work is ongoing.

Additionally, NWEA conducted an alignment study focused on the correspondence between existing MAP Growth items and the Georgia state content standards. Local educators reviewed items that currently exist in the MAP Growth item pool and evaluated their alignment to the Georgia standards. For the items that did not align but came close, revisions were suggested. The TAC suggested that using the preexisting items will help with their development efforts and could be beneficial for scaling.

The COVID-19 pandemic has impacted some of NWEA's assessment development activities. Most meetings and interactions this calendar year have been conducted virtually, as will the content and bias review meetings scheduled for July 2020. Some activities have been postponed, including phase two of the Achievement Level Descriptors utility study, a

comparability presentation to superintendents, and the field test that was scheduled for Spring 2021.

The field test and adaptive test design plans are still under development. Following some internal discussions, NWEA is considering online calibration, which targets item parameter precision rather than sample size. Specifically, a standard error of measurement criterion determines the stopping rule for field testing each item. The TAC agreed that this is an approach worth exploring and suggested that the GMAP Partnership also consider how this plan ensures representation of the consortium's full student ability distribution.

The TAC suggested that NWEA use existing parameter estimates from items in the MAP item bank. If these items' parameters are fixed during calibration, only new and revised items need to have item parameters estimated. The TAC noted that another strategy to explore is using existing item parameter estimates as Bayesian priors.

The GMAP Partnership and NWEA are also discussing whether the test that is being developed will ultimately be item adaptive or multistage. Item adaptive testing becomes challenging, of course, with language arts assessments that are composed of passage-based blocks of items. The TAC expressed concern regarding the alignment of the depth and range of knowledge within a given subject or domain. NWEA shared that, from a design standpoint, their item development plan and item specifications ensure that the breadth and depth of each assessable standard is represented. The TAC suggested that their alignment concern could also be addressed by using staged adaptive testing, and that alignment could be evaluated quantitatively by including it as a criterion in NWEA's simulation studies.

The GMAP Partnership next discussed the field test plan — in particular, the sample size needed to estimate item parameters for the operational item bank. If the sample size needs to increase, there are additional districts that the GMAP Consortium may be able to recruit to participate in the field test who are not already MAP Growth users. The TAC reminded the consortium to balance sample-size needs against administration logistics and student motivation; item parameter estimates from standalone field test items are usually less accurate and precise than embedded field test items. However, the TAC noted that limited student motivation could be less of a problem if the assessment generates useful information that NWEA could provide back to the participating schools. The TAC also suggested that in order to get a large enough sample, a MAP Growth test — with embedded items from the through-year assessments — could be administered free of charge across the state. Through-year field test items could be embedded into the nationwide MAP growth test; NWEA would want to confirm that parameter invariance holds (i.e., that the item parameters estimated via national data would be essentially unchanged if they were

estimated via state-level data), but given the state's diversity and wide range of student achievement, parameter invariance is unlikely to be a major concern.

The GMAP Partnership also noted that item development has been informed by range achievement level descriptors (ALDs) that are somewhat different from the Georgia Milestones ALDs. The Partnership was asked whether these new range ALDs would preclude achievement-level comparability between Milestones and the GMAP through-year system (achievement-level comparability is required if a consortium intends for its students to take its innovative assessments in lieu of Milestones). GMAP responded that its range ALDs simply elaborate upon the Milestones ALDs and are used in conjunction with the item specifications to inform the item-writing process. It will be important to check in on this issue again in future technical assistance sessions or TAC meetings, since achievement-level comparability (and, presumably, ALD similarity) is required for innovative assessments under IADA.

The TAC also inquired about how data from each testing event would be used in accountability, noting that in order to be valid, a proficiency calculation must be based on results across the entire test blueprint/standards. The GMAP Partnership shared that students will take every test event in fall, winter, and spring regardless of proficiency level. Test events will be designed to have content constraints that are consistent across time. The TAC suggested that if students know they are proficient based on the winter test, they may not have the same motivation to perform well when they test in the spring. The TAC recommends that NWEA think more about the student-level reporting and how student motivation might be impacted by the through-year design. One possible approach would be to provide districts and teachers with specific diagnostic information on how students are performing on given standards.

GMAP Demographics and Achievement Metrics

Description

NWEA presented a demographic summary of students in the GMAP consortium, along with their corresponding achievement on Georgia Milestones assessments. When compared with the state of Georgia, Hispanic, African American, and economically disadvantaged students are overrepresnted in the GMAP consortium. The TAC was asked to provide input on ensuring representation during field testing in accordance with the IADA and to suggest strategies to ensure representation is maintained for the calibration of the through-year scale as the consortium grows during field testing years.

TAC Discussion and Recommendations

NWEA's presentation included a review of the member districts, the number of students tested in each grade and district, a comparison of MAP districts' demographics with those of

the state and non-MAP districts, and student achievement levels in English/Language Arts (ELA), math, and science.

Since Hispanic, African American, and economically disadvantaged students are overrepresented in GMAP districts (compared to the rest of the state), the TAC was asked to weigh in on two issues: (1) how the consortium should sample students to ensure representation and (2) whether this representation needs to be of the GMAP member districts or of the state. The TAC shared that the intent of IADA is to include demographically diverse districts. The GMAP Consortium can use a representative sample of the member districts but should clarify that, as their district membership grows, they will move closer to the end goal — statewide representativeness. The TAC suggested that if GMAP selects a stratified sample of their districts to be representative of the state, the Partnership could then examine the demographic differences between that sample and the full GMAP Partnership membership. Over time, as the Partnership grows, those differences should narrow.

NWEA followed up with a question about planning for test-taker population change over time: How should the Partnership plan for and then leverage or mitigate major shifts in demographics with the addition of new member districts? The TAC suggested that the approach depends on the confidence NWEA has in the original scale from the first year of field testing. If NWEA is not confident that the scale is stable, then the addition of new districts can be an opportunity to add item response data and improve the scale. The TAC also suggested that NWEA consider recalibrating the scale every year, with the final year producing the final scale. The TAC also emphasized that the stability of the scale would be more severely impacted by interruptions to the school year due to COVID-19 than from shifts in demographics.

TEST SECURITY

Description

NWEA described their test security practices for the GMAP through-year assessment to the TAC. The presentation discussed test security standards through test design and development to test administration. The presentation detailed test security montioring and detection processes. The TAC was asked to provide feedback on the procedures and practices that were presented.

TAC Discussion and Recommendations

NWEA presented on their test security standards and procedures for maintaining security before, during, and after test administrations. NWEA shared that they received Caveon's *Seal of Excellence* after undergoing a test security audit. This certification recognizes strong test security practices and policies. Caveon worked with NWEA to develop a comprehensive test

security plan which NWEA shared with the TAC. For the through-year solution planned in Georgia, NWEA does not currently foresee the need for deviation away from its standard operating procedures for secure testing.

The TAC requested data that might provide evidence of the effectiveness of the procedures in place on the GMAP through-year assessment. Relevant data might include the number of testing irregularities that are reported, the extent to which test administrators are following the test administration manuals, the findings from incident investigations, and the number of times items have been compromised on a web search.

The TAC affirmed that the procedures in place are quite strong, particularly under normal testing conditions. Given that schools are exploring alternative plans for the 2020–2021 school year (e.g., virtual learning), the TAC recommended that the GMAP Partnership explore how test security may need to be relaxed under abnormal circumstances. At the next TAC meeting, there may be further discussion about what validity or security sacrifices may be necessary in order to record scores and provide feedback to schools.

The TAC offered suggestions on how to communicate test security rules to students, particularly because the assessment has an extended testing window. In many cases, cheating occurs because students do not realize what the rules are and which behaviors (e.g., conversationally sharing answers, discussing passages) are not appropriate. The TAC suggested this problem could be mitigated by having students sign a waiver affirming that they understand the rules.

The TAC also inquired about prior exposure of test items over an extended period of time. NWEA responded that, because there is a large item bank, students should not see the same items over multiple testing events. NWEA also conducts statistical checks on the items to flag irregularities (for example, item parameter estimates drifting over time due to exposure). NWEA is also exploring options for dividing the item pool into “less exposed” and “more exposed” subgroups of items.

Protecting Student Data Privacy

Description

NWEA described their data privacy protocols, information security system, and audit and compliance procedures for maintaining the security of student data. The TAC was asked to provide feedback on their proposed procedures.

TAC Discussion and Recommendations

NWEA shared that their Information Assurance department oversees activities that support privacy, information security, compliance, cybersecurity risk management, and test security. The TAC suggested that the GMAP Partnership should plan to conduct risk-management

activities along with the Georgia Department of Education (GaDOE) in the future. For example, a review of procedures, and roles and responsibilities should be conducted. NWEA shared that they already have some security and compliance practices in place when they work with an education agency such as GaDOE.

Maintaining Data Integrity

Description

NWEA described their procedures for ensuring data quality, along with their standard operating procedures for data management before data is transferred to state reporting systems. The TAC was asked to provide feedback on the proposed procedures and practices to maintain data integrity.

TAC Discussion and Recommendations

NWEA explained that they have a deep commitment to ensuring quality through each step of their process and guided the TAC through their data classification information, data definition standards, and the dimensions of data quality that they emphasize and track in their work. Additionally, NWEA presented information on their data management process options they typically use with their clients.

The TAC asked for additional information about NWEA's rostering process for schools and districts. NWEA shared that they have many options the GMAP Consortium can use. One option would be a single-file system with the state; alternatively, NWEA can allow local education agencies to upload their data individually. GaDOE shared that for the summative assessment system, they do not get frequent data updates from their districts. GaDOE suggested that for a through-year assessment system, it would be best to work with districts directly to ensure rostering information is up-to-date at the time of test administration. The TAC also reminded NWEA that they have responsibilities on both ends of the rostering system — in getting student data input into the system by districts, and then also reporting that data for the state.

NWEA also discussed the regular statistical key checks that they are currently conducting for their summative assessment clients. NWEA expects that they will need to make some modifications for the Georgia through-year model. The TAC asked how easy it is to look up the statistical specifications of an item as it makes its way through field testing. NWEA shared that they are updating their item management system and anticipate that they will be able to view item parameter estimates and related statistics across time once the through-year item field testing begins.

The TAC also asked how NWEA's standard demographic categories align with federal requirements. NWEA shared that they will make sure the groups represented in Georgia and

that are required for federal reporting will be included in their standard operating procedures.

The TAC recommended that NWEA also consider planning for unexpected changes over time. As the test is scaled up, there might be instances where districts have unexpected increases or decreases in scores. The TAC recommended that the data system be set up in a way that the data needed to investigate these unexpected changes are easy to access. For example, demographic changes in a particular district, individual student performance data over time, and district performance over time might need to be accessed. Additionally, in order to account for possible changes in scores that could be attributed to changes in curriculum and students' opportunity to learn, the GMAP Partnership should consider regularly asking districts if they are implementing any new initiatives, so that there is a starting point for hypotheses.

NEXT STEPS

Future TAC Meetings

During the debrief with GaDOE and WestEd, the TAC requested the following information during future TAC meetings:

- Additional details on how results will be presented to stakeholders (e.g., mockups of individual student reports)
- Updates on COVID-19's impact on the Partnership's plans and activities, including how alternative instructional scheduling may impact the data they plan to collect in 2020–2021
- Results of any studies that have been conducted, preferably with summaries that emphasize how the study findings can be used as evidence to support decisions about the through-year assessment program. The TAC assumes these studies will include NWEA's analysis of item-level alignment data.
- Plans for scaling as the consortium membership grows
- More information about the shadow CAT approach and the benefits of implementing it

The TAC also recommended that each consortium discuss the following topics in future TAC meetings:

- Comparability within the assessment system (e.g., across forms and testing occasions in a through-year or otherwise distributed test design)
- Updates on any independent alignment studies that have been conducted
- Reporting

APPENDIX 2: INNOVATIVE ASSESSMENT PILOT APPLICATION ASSURANCES

Alignment

- Aligns with Georgia's academic content standards (breadth and depth of those standards for all grade-levels and content areas or courses assessed)
- Identifies which students are not making progress toward Georgia's academic content standards
- Produces results that are comparable to the Georgia Milestones assessments (include methods in the narrative or as attached evidence)

Technical Quality

- Works with expert(s) (external partner or in-house) to ensure technical quality, validity, reliability, and psychometric soundness of the innovative assessment
- Establishes validity and reliability evidence consistent with nationally recognized testing standards
- Assesses student achievement based on state academic content standards in terms of content and cognitive processes, including higher-order thinking skills, and adequately measures student performance across the full performance continuum
- Produces individual and aggregate reports that allow parents, educators, and school leaders to understand and address the specific needs of students
- Provides reports in an easily understandable and timely manner to students, parents, educators, and school leaders
- Developed, to the extent practicable, consistent with the principles of universal design for learning

Accommodations

- Appropriate accommodations will be provided for students with disabilities as defined via their IEP or IAP (provide list of available accommodations as an attachment)
- Appropriate accommodations will be provided for English Learners as defined via their EL/TPC (provide list of available accommodations as an attachment)

Security

- Develops and implements policies and procedures to ensure standardized test administration (i.e., test coordinator manuals, test administration manuals, accommodations manuals, test preparation materials for students and parents, and/or other key documents provided to schools and teachers that address standardized test administration and any accessibility tools and features available for the assessments)
- Delivers training for educators and school leaders to ensure a standardized test administration
- Develops and implements a monitoring process to ensure standardized test administration
- Develops and implements policies and procedures to prevent test irregularities and ensure the integrity of test results
- Develops and implements policies and procedures to protect the integrity and confidentiality of test materials, test-related data, and personally identifiable information

Stakeholder Engagement

- Develops assessment in collaboration with stakeholders representing the interests of students with disabilities, English learners, and other vulnerable populations; teachers, principals, and other school leaders; parents; and civil rights organizations
- Develops capacity for educators and school and district leaders to implement the assessment, interpret results and communicate with stakeholders

Accountability

- Produces a single, summative score for every student
- Produces a comparable growth measurement that can be used for the Progress CCRPI component
- Produces a comparable achievement measurement that can be used for the Content Mastery and Closing Gaps CCRPI components (alignment to Beginning, Developing, Proficient, and Distinguished Learner achievement levels)
- Produces a comparable literacy (Lexile) measurement that can be used for the Readiness CCRPI component
- Produces subgroup results consistent with federal accountability and reporting requirements (e.g., race/ethnicity, gender, English Learners, students with disabilities, migrant, homeless, foster, parent on active military duty)