Georgia Innovative Assessment Pilot Program

TECHNICAL ASSISTANCE ANNUAL REPORT

Sonya Powers

Assessment Research & Innovation @WestEd | csaa.wested.org

January 2022

TABLE OF CONTENTS

GEORGIA INNOVATIVE ASSESSMENT PILOT PROGRAM

TECHNICAL ASSISTANCE ANNUAL REPORT

INTRODUCTION

The purpose of this Technical Assistance Annual Report is to summarize how the technical assistance needs of Georgia's Innovative Assessment Pilot Program (IAPP) consortia have been addressed through meetings with a Technical Advisory Committee (TAC) and meetings with WestEd, Georgia's IAPP technical assistance provider, during the second year of implementation. Lessons learned and recommendations for future pilot program activities are also included.

During the first year of implementation, as described in the Year 1 IAPP Technical Assistance Annual Report, a number of key themes emerged:

- delays due to COVID-19 and impacts to the IAPP timelines,
- challenges of comparability and assessment for accountability,
- resource challenges associated with building and scaling new assessments, and
- benefits and limitations of an assessment competition.

These themes have carried forward into Year 2. In fact, as disruptive as COVID-19 was during the 2019-20 school year, 2020-21 was in many respects worse. Although most schools offered in-person instruction in Fall 2020, COVID-19 cases and rolling quarantines resulted in continued disruptions to education. Rather than impacting the last two or three months of school, the pandemic resulted in profound changes to education for the entire school year. States, including Georgia, again sought waivers from the federal government for statewide accountability testing in spring 2021. Although the federal government did not permit testing to be cancelled for a second year, test results were not used for federal accountability. Nevertheless, given concerns about health, safety, and instructional time, testing may have been seen as a lower priority: student participation rates in spring 2021 for the Georgia Milestones assessments were noticeably lower than usual, dropping from an average of 99% in 2019 to a range of 59% to 78% in 2021, depending on grade and subject. Given the havoc the pandemic has wreaked within and far beyond the education system, Georgia's IAPP has also faced delays and slow progress. Neverthless, the two consortia—the Georgia MAP Partnership and the Putnam Consortium—have continued to move forward with developing their assessment programs, while pivoting to serve their partner school districts during this challenging time.

In this Year 2 report we describe the areas where the two consortia have made progress, the impact of pandemic delays on each consortium's timelines, and the process of defining the evaluation criteria to determine whether the consortia assessments may be used in lieu of the current statewide

assessment system. We also summarize the technical assistance provided by WestEd and the TAC. The psychometric issues highlighted in the narrative are described in greater depth in Appendix 1, which includes four TAC reports—one for each consortium summarizing the TAC meetings held in December 2020 and July 2021.

Program Requirements and Technical Assistance Priorities

Georgia's IAPP was authorized under Georgia Senate Bill 362 and the United States Department of Education Innovative Assessment Demonstration Authority (IADA). Two groups of school districts—the Putnam Consortium (Putnam) and the Georgia MAP Assessment Partnership (GMAP)—were granted the authority to develop new accountability assessments. Districts participating in the GMAP and the Putnum consortia can administer a new assessment program (either the Georgia MAP assessment in the GMAP consortium or the Navvy system of assessments in Putnam) in place of the state's summative Georgia Milestones tests once the new assessments have demonstrated comparability to Georgia Milestones and received approval from the state. The original timeline for the consortia to demonstrate comparability was a five-year period, beginning in fall 2019 and completing in summer 2024. It may be possible to receive a two-year extension from the federal government, which would allow the pilot to continue through summer of 2026.

To support the Putnam and GMAP consortia, the Georgia Department of Education (GaDOE) contracted with WestEd to provide technical assistance to both consortia. Technical assistance is provided through two primary mechanisms: 1) WestEd meetings with the consortia to discuss the IAPP goals, project roadblocks, and psychometric considerations, and 2) twice-yearly technical advisory committee (TAC) meetings facilitated by WestEd where the consortia can get assessment advice from industry experts. One important outcome of the Year 2 technical assistance was the formalization of Comparability Guidelines. This section will summarize the WestEd-consortia meetings, the development of the Comparability Guidelines, and the TAC meetings.

WestEd-Consortia Technical Assistance Meetings

Due to budget cuts within GaDOE, funds for WestEd staff time to provide direct technical assistance were significantly reduced. During Year 1, 114 hours of WestEd staff time were available to Putnam and GMAP, compared to only 12 in Year 2. Despite this reduction, the consortia did not use all of the hours. GMAP used 8 while Putnam used only 1. One possible explanation for the lack of use of the technical assistance is that planned data analysis was not possible after spring 2020 due to pandemic testing cancellations. Thus, comparability analyses and related psychometric considerations were put on hold. During the meetings with the two consortia during Year 2, WestEd worked with the consortia on preparations and topic selection for upcoming TAC meetings, comparability and statewide accountability readiness, and alignment studies. WestEd also served as a liaison between the

consortia and GaDOE when questions about Georgia Milestones policies and documentation or comparability requirements arose.

Nevertheless, better use could be made of WestEd's technical assistance, which is available at no cost to the consortia. For example, validity and comparability research plans could be discussed, analysis specifications could be reviewed, and several aspects of comparability that are not reliant on data could have been explored (e.g., test administration and security, stakeholder engagement). WestEd will continue to encourage the consortia to make active use of the technical assistance hours, identifying potential topics to discuss, and leveraging some of the hours for review of comparability documentation.

During Year 2, WestEd used the remaining technical assistance hours that had not been used by the consortia to develop a Comparability Guidelines document (see Appendix 3 for the full document; more description can be found in the section that follows).

Comparability Guidelines

It is an IADA requirement that comparability be established for a new assessment before it can be used in lieu of the state's existing accountability assessment. Thus, comparability has always been top-of-mind for Georgia's two IAPP consortia. The IADA comparability requirement is that students receive equivalent achievement level classifications regardless of the assessment they take. In other words, a student classified as proficient on Georgia Milestones should also be classified as proficient by Navvy or GMAP. However, the IADA statistical comparability requirement is a small part of the comparability evidence that the consortia must provide to the Georgia Department of Education for evaluation. As part of their IADA applications, the consortia also committed to other requirements, such as making accommodations available for English learners and students with disabilities to allow for their participation in the consortia assessments at the same rates that they would participate in state assessments (see assurances in Appendix 2).

During Year 1 and the beginning of Year 2, the consortia and TAC discussed comparability and the associated requirements for providing valid and reliable data to be used in Georgia's state accountability system. Throughout Year 2, the following questions were revisited:

- What evidence would the TAC deem sufficient for performance level comparability?
- What were the specific criteria that the consortia would be held to when their assessment programs were evaluated?

To help address these questions, Comparability Guidelines were documented to serve as a comprehensive checklist, similar to the peer review templates that states must submit to the U.S. Department of Education.[1] The Comparability Guidelines build on the original assurances, making the

[1] https://www2.ed.gov/admins/lead/account/saa/assessmentpeerreviewsubmissionindexacdemic.doc

requirements more concrete, and providing examples of the types of evidence that address each of the requirements. WestEd drafted the Comparability Guidelines, they were reviewed by both GaDOE and the IAPP TAC, feedback was incorporated, and the final set was approved and provided to the two consortia in July 2021.

As noted in the Year 1 report, the comparability criteria related to achievement level classifications is not an unattainable bar for the consortia to meet. However, other requirements that existing state assessments have to meet for federal and state accountability purposes (e.g., test security, accommodations for students with disabilities and English learners) significantly increase the demands on the consortia assessments. The Comparability Guidelines document describes six different categories with a total of over 30 separate criteria for which consortia assessments must provide evidence to ensure that they can support the same high-stakes decisions that are currently made on the basis of Georgia Milestones scores. Specifically, the state uses student scores on Georgia Milestones for grade retention and promotion decisions, as part of course grades in high school, in teacher and leader evaluations, and as a key component of its College and Career Ready Performance Index (CCRPI) accountability metrics. Consortia assessments must therefore meet a high bar for quality, accessibility, security, and other aspects of their assessments.

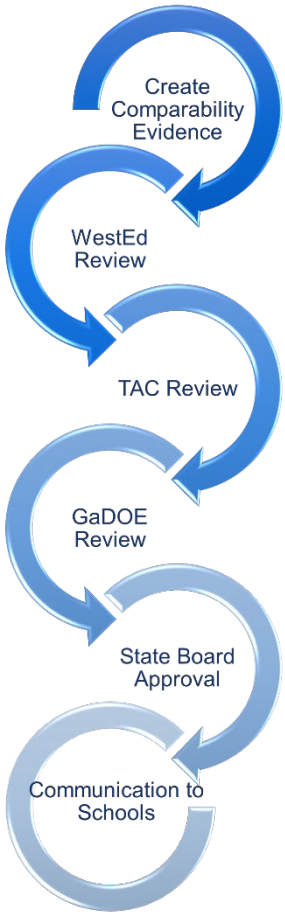
One concern raised in the Year 1 annual report was that the timeline might already be too short for the consortia to assemble all of the necessary comparability evidence, have it reviewed by the TAC and GaDOE, and be approved for use in lieu of Georgia Milestones within the five-year project timeline. As shown in Figure 1, it is likely that the first operational administration could not take place until 2024-25, beyond the current five-year pilot program timeline.

Figure 1. Current IAPP Timeline

Given the disrupted 2020-21 school year, it is even more likely that Year 6 might be the first year of implementation of GMAP or Navvy in lieu of Georgia Milestones, unless comparability can be fully established using data from 2021-22. Furthermore, both consortia have planned to establish comparability for English Language Arts (ELA) and mathematics assessments first, with science following by one year. Thus, Year 7 might be the first year of implementation of GMAP or Navvy in place of Georgia Milestones for science. Additionally, Grade 8 social studies and U.S.History were not part of the original plans submitted by the consortia in their IADA applications, yet they are part of the current statewide assessment system and will also need to be provided by the consortia in the future, meaning implementation of the full suite of Georgia Milestones-comparable assessments is likely at least two years beyond the original project timeline.

Another impact to the timeline is that the evidence submitted to document comparability will need to go through a series of review steps (see Figure 2). First, the consortia will provide information to WestEd, who will review for completeness and then route it to TAC members for review once it is deemed ready. The TAC will then review the documentation, provide feedback, and if necessary, review revisions. Once the TAC approves the documentation as complete and adequately supporting comparability, a GaDOE state panel will review it. Once GaDOE signs off, the State Board of Education will review for final approval. Should assessments be approved, consortium districts will be notified that the consortium assessment can be used in place of Georgia Milestones and their accountability evaluations. Because both Navvy and GMAP are through-year assessments, schools, parents, and students will also need to be notified of a change in assessment used for accountability prior to the start of the school year because the first administrations of the through-year assessments could start soon after the school year begins.

Figure 2. Comparability Review Process



The multi-step nature of this review process will take some time. To make the process more efficient, WestEd is working with the consortia to stagger the flow of information. Nonetheless, it is critical that the evidence be thoroughly reviewed and strengthened as needed through the process, as some of the same types of evidence would ultimately be required for federal peer review if one of the consortia assessments becomes the statewide assessment system in Georgia.

Biannual TAC Meetings

WestEd planned and hosted two TAC meetings during Year 2 of the Georgia IAPP. Each consortium met with the TAC for one day at each meeting. Participant districts, their test development partners, WestEd, GaDOE, the Governor's Office of Student Achievement (GOSA), and the TAC's expert advisors took part in the TAC meetings. The meetings, convened virtually, took place December 14–15, 2020 and July 7–8, 2021. The IAPP TAC includes the following assessment policy and measurement experts:

- Wayne Camara, Distinguished Scientist for Measurement Innovation, Law School Admissions Council
- Gregory Cizek, Professor of Educational Measurement and Evalution, School of Education, University of North Carolina at Chapel Hill

- Stuart Kahl, Senior Technical Consultant/Advisor in Assessment, Kahl Balanced Assessment Practices
- Lillian Pace, Senior Director of National Policy, KnowledgeWorks
- Stanley Rabinowitz, Senior Technical Advisor, Pearson
- Steven Sireci, President, Sireci Psychometric Services

WestEd facilitated the TAC meetings and worked with the consortia to create an agenda of topics on which TAC feedback and advice was desired. During the July meeting, WestEd presented the Comparability Guidelines, and approximately half of the meeting was dedicated to updating the TAC and the consortia about the document as well as providing time for questions and answers. Both before and after the TAC meeting, members of the TAC provided feedback on the Comparability Guidelines. Once the feedback was incorporated, the TAC approved the final version.

During the biannual meetings, the TAC provided advice about both technical and pragmatic aspects of each consortium's assessments. They also helped to identify issues that the consortia may not have considered, but which could become very important issues to address. For example, the TAC noted that both consortia would need to determine how to handle a student who moves into a district or state midway through the year. For Navvy, administering a separate assessment for every standard in such cases may not be feasible, and an alternative will be needed. Likewise, GMAP must consider how to assess students who were not in the district during fall and winter administrations if those administrations would typically contribute to students' summative scores.

The specific process for calculating summative scores has yet to be determined by either consortium. During Year 2 meetings, TAC members pushed both consortia to finalize their approach, given that it is fundamental to establishing comparability and must be decided in order to complete field test analyses in spring 2022. The TAC also encouraged the consortia to think about the definition of the summative score and what it reflects in terms of how learning is measured in its calculation. For example, should the summative score be a summation of scores that reflect student content mastery immediately after instruction or should it reflect content knowledge retained at the end of the school year? The way learning is defined by the consortia and described through the summative score may or may not be consistent with the way it is defined and described through summative scores on Georgia Milestone. Thus, the definition of content mastery may not be strictly comparable, and TAC members advised that differences be carefully considered and justified.

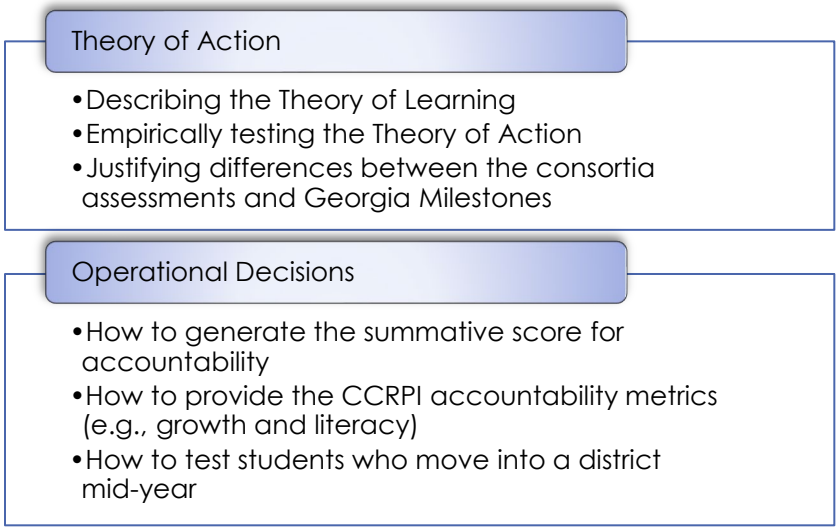
In fact, a consistent theme throughout the TAC meetings was that the consortia should critically evaluate differences between their assessment solution and the current state content standards and assessments. The differences should not only be justified based on a consortium's theory of action (e.g., greater instructional or diagnostic value), but these theories should be empirically tested to provide evidence that differences are leading to improvements.

The TAC also cautioned about using 2020-21 results for comparability analyses given concerns about opportunity to learn and motivation during an administration that did not count for federal

accountability. The TAC also noted that one of the most important considerations for any analysis is the representativeness of the consortium's participants in comparison to the state's demographic and achievement profile. Without representativeness, results may not be generalizable. Thus, the consortia should evaluate representativeness each year as participating districts join and leave.

Finally, the Comparability Guidelines presented in July clarified how the consortia assessments would need to support calculation of the state's CCRPI accountability metrics, and TAC members noted that the consortia will need to explore options and determine how they will provide similar metrics for state accountability. Figure 3 provides a summary of the TAC feedback from the two meetings held during Year 2.

Figure 3. Summary of 2020-21 (Year 2) TAC Feedback



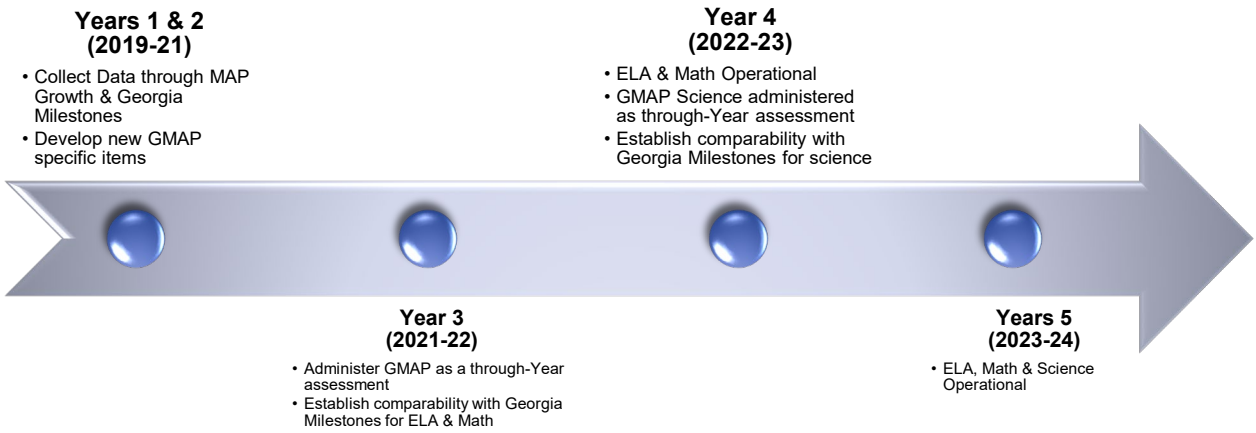
PROGRESS TOWARD FULL IMPLEMENTATION

GMAP is based on NWEA's MAP Growth assessment system, which was used in some Georgia school districts prior to IAPP. Likewise, Navvy ELA and mathematics assessments have been administered in the Putnam school district since 2017. Thus, both consortia began the IAPP by leveraging assessments that were used in Georgia prior to the pilot. NWEA and Navvy have existing item pools, established test designs, and psychometric modeling decisions that provided a basis upon which to build out their assessment solutions. Nevertheless, the pandemic has impacted the original timelines proposed in Georgia's IADA application, pushing back some benchmarks by at least a year.

Figure 4 shows the original GMAP timeline. GMAP had dedicated time in the first two years to understanding the alignment of MAP Growth assessments to the Georgia Standards of Excellence and developing new items for GMAP to better align to the Georgia standards. This work has moved forward despite the pandemic, and thus, GMAP's timeline has not been impacted as greatly as it might

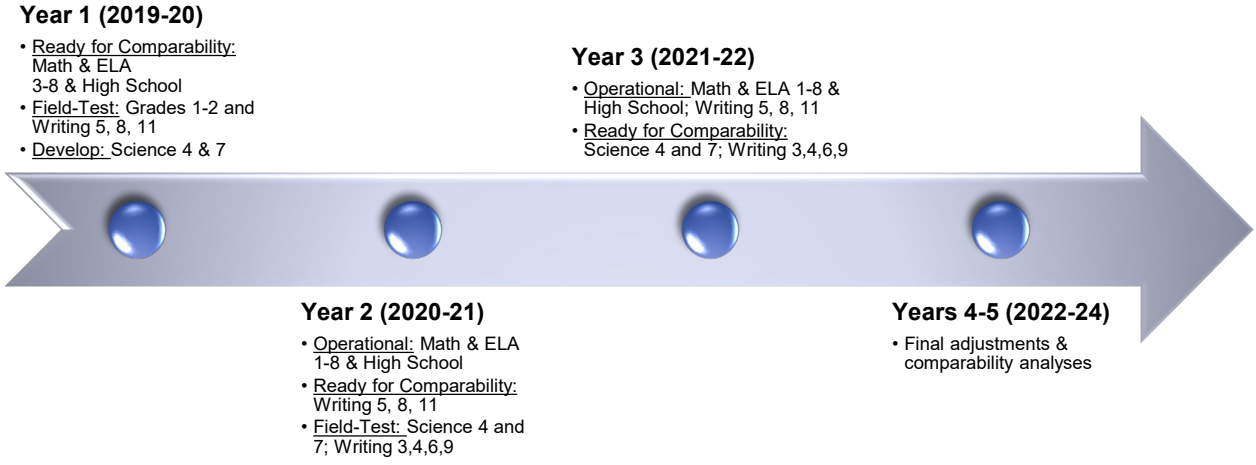
have been. However, data collections planned for spring 2020 and the 2020-21 school year were delayed. Thus, the first time GMAP items will be administered will be spring 2022, and the first time GMAP will be administered as a through-year assessment will be delayed from Year 3 (2021-2022) of the project to Year 4 (2022-2023).

Figure 4. Original GMAP Timeline



By contrast, Putnam's original timeline front-loaded many activities, using the final two years to make necessary adjustments to the assessment system and scaling to additional districts (see Figure 5).

Figure 5. Original Putnam Timeline

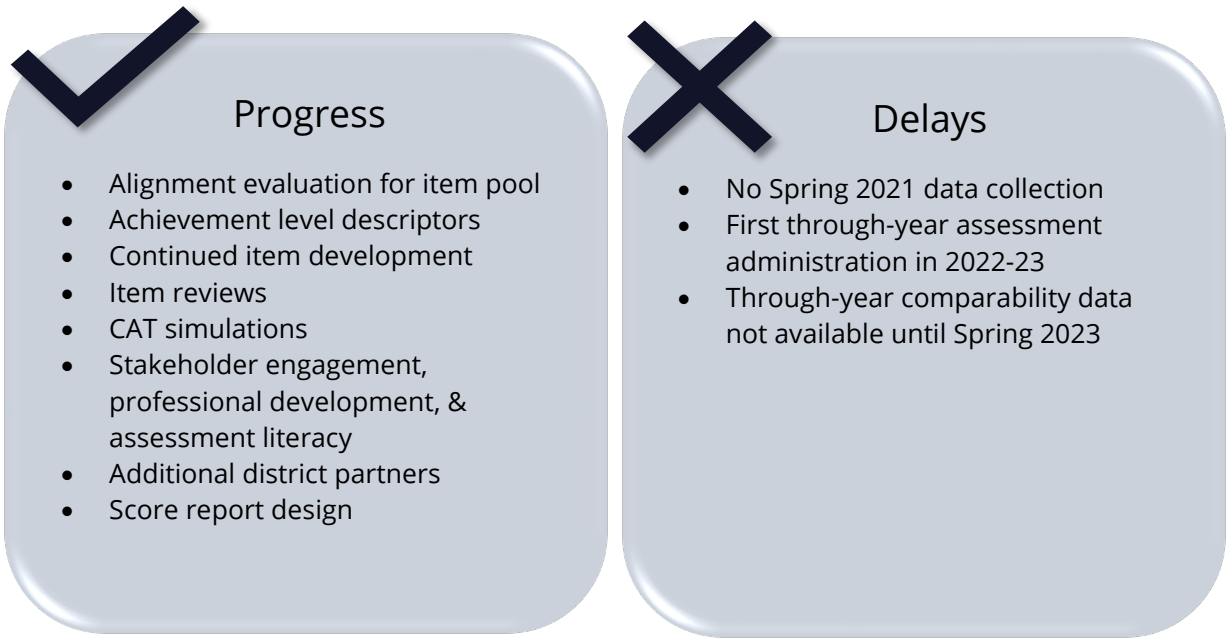


The Putnam consortium's priority was to establish comparability quickly and obtain approval to use Navvy instead of Georgia Milestones as soon as possible so that consortium members would not need to continue using both assessments. However, these plans were interrupted when the Georgia Milestones was not administered in Spring 2020 and adminstration of Navvy was likewise interrupted during the 2020 spring semester. Continued disruptions in 2020-21 pose a challenge for establishing comparability in Year 2 of the pilot. Using Georgia Milestones Spring 2021 results and Navvy 2020–21 results for comparability may be difficult due to pandemic-related disruptions which impacted data completeness and quality for both assessments. Thus, the 2021-22 school year is the first school year where statistical comparability can be thoroughly evaluated, assuming all goes to plan. Item development work for Science has also been delayed. Putnam's plan was always to stagger the rollout of science but the rollout will likely be slower given the delays. Although it appears likely that the benchmarks in Putnam's original timeline will all shift back two years, Putnam was able to collect data during the first two years of the program, allowing them to conduct preliminary analyses on item performance. The consortium showed rates of standard mastery and average item discrimination values for Grade 4 math during July 2021 TAC meeting. Results indicated that there were differences in the proportion of participating students mastering each standard; average item descrimination values were all above 0.3, indicating that many Navvy Grade 4 math items appear to perform well enough to be considered for an operational statewide assessment.

The timelines shown earlier illustrate the rollout of each consortium's assessment and the target dates by which they could be used in lieu of Georgia Milestones. The figures do not show all the other activities the consortia completed during the first two years of the pilot. Many activities were able to continue virtually such that they did not depend on having teachers and students in school buildings.

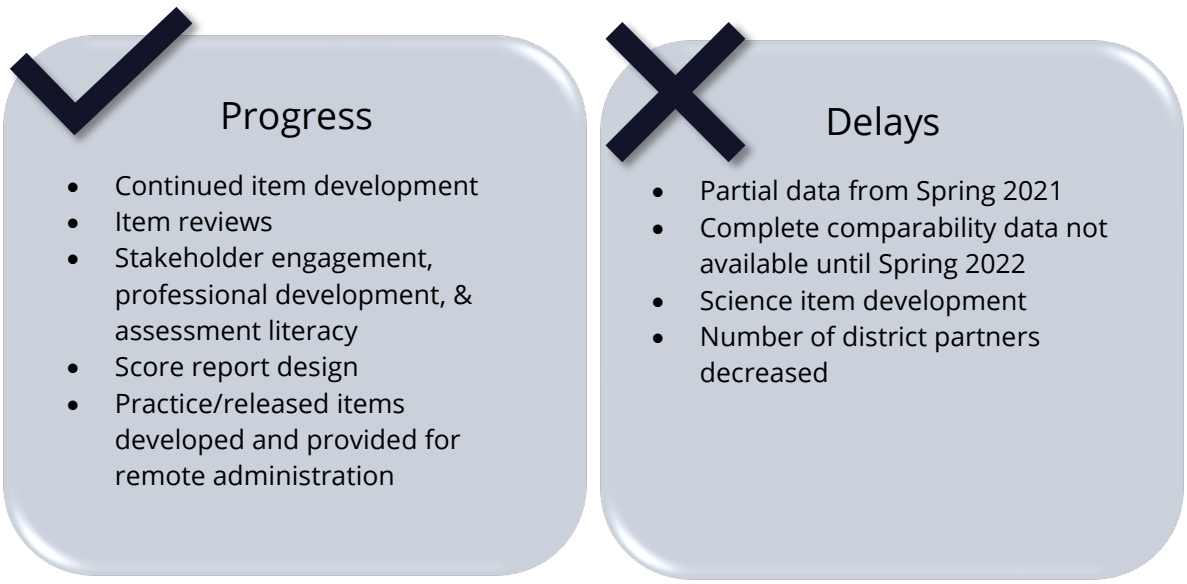
For example, GMAP conducted a MAP-to-Georgia content standards alignment study to identify gaps in alignment. NWEA identified item banks that could support the GMAP assessment, and created and began implementing an item development process to create new items to assess Georgia standards not covered by existing items. GMAP also involved educators in a review of achievement level descriptors based on Georgia's existing achievement level descriptors. Achievement level descriptors were also incorporated into the item development plan so that item writers would have guidance to support development of items aligned to the Georgia standards that also span the range of student proficiency. NWEA also conducted item reviews, including bias and sensitivity reviews, virtually. NWEA was also able to continue refining their computerized adaptive testing (CAT) algorithm via simulation studies to better understand how many items are needed to yield accurate and reliable student scores that appropriately align to the breadth and depth of Georgia's content standards. GMAP continued to work with stakeholder groups, providing professional development services around assessment literacy, as well as getting score user feedback on new score reports in development for the GMAP assessments. Finally, the GMAP consortium maintained its partner districts and added 11 additional districts to its membership, including 5 that participated in Year 2 and 6 more that have signed on for Year 3. Figure 6 provides an overview of implementation progress for the GMAP consortium during Year 2.

Figure 6. Overview of GMAP Implementation Progress during Year 2



The Putnam consortium was also able to continue item development efforts during the first two years of the pilot. In fact, they embarked on an ambitious project to develop a set of practice items that could be administered to students remotely. These items were developed to provide educators with an understanding of the content covered by Navvy and the level of difficulty of the items. For test security reasons, the secure Navvy items are not available to teachers and are not available for remote administration. Practice items helped teachers assess students and continue to use Navvy to support instruction. Navvy also continued stakeholder engagement during Year 2, continued to provide professional development to district partners, and built out and refined their student level and aggregate score reports, based on feedback from score users. Although schooling disruptions resulted in less-than-complete Navvy data for most Putnam consortium districts, some participating districts were able to implement many of the Navvy assessments. Data analysis is ongoing to support item reviews and begin to investigate comparability with Georgia Milestones. District membership in the Putnam consortium increased for Year 2 of the pilot, but some districts have not agreed to continue participation for Year 3. Nevertheless, Putnam has retained committed local supporters in the consortium. Figure 7 provides an overview of implementation progress for the Putnam consortium during Year 2.

Figure 7. Overview of Putnam Consortium Implementation Progress During Year 2



Both consortia solicited feedback from WestEd and the TAC on the technical aspects of their assessment systems. Many considerations were discussed beyond statistical comparability including accommodations, reporting, and test security. For example, both consortia asked whether accommodations could be phased in over time. The TAC understood that low-incidence accommodations (e.g., Braille) might not be ready for the field test administrations, but all accommodations needed to provide students with appropriate access to the test content should be available as soon as possible, and definitely before the assessment would be used for accountability purposes. Both consortia have also been working on score report refinements during the pandemic and Putnam presented some dashboard displays during TAC meetings. The TAC has expressed interest in discussing score reports in more detail and getting more specific information about how stakeholders have been engaged to ensure the usefulness of score report information. Test security, which is an element of the comparability evidence that the consortia must provide, has also been discussed at a high level with the TAC. The TAC advised that rigorous test security procedures are needed for any administration that contributes to a student's summative score. If the consortia wish to include through-year assessment opportunities that do not contribute to a student's score, less rigorous security procedures might be reasonable so long as the item pool for assessments that contribute to the summative score is kept separate.

With high hopes that spring of 2022 will provide complete assessment data, the consortia are working toward submitting comparability evidence. Thus, timelines and procedures for submitting comparability evidence for review have been the focus of discussion. Both consortia desire an efficient process so that member districts can stop using both innovative assessments and Georgia Milestones as soon as possible. As mentioned previously, the consortia (particularly Putnam) requested very little

technical assistance from WestEd during Year 2. This technical assistance can serve as evidence in support of the first criterion in the Comparability Guidelines related to technical quality, which asks:

> *Have you worked with experts to ensure technical quality, validity, reliability, and psychometric soundness of the innovative assessment?*

WestEd will continue to work with the consortia in Year 3 and encourage increased use of technical assistance support, and particularly to press for timely submission of high-quality materials to the TAC.

Although the consortia have made strides given the constraints of the past two years, the IAPP period is reaching the halfway point of the original 5-year timeline. The TAC has expressed concern about the number of decisions, analyses, and results still needed to ready the consortia for administration in lieu of Georgia Milestones within the pilot period. Critical decisions that need to be made include determining how a summative score is calculated, determining how growth and literacy measures are calculated for CCRPI, and developing an assessment plan for students who are only in the consortia for part of the year. Analyses and results include statistical comparability, reliability and validity calculations, and independent alignment studies.

More generally, the TAC also noted a desire for more detailed TAC materials (i.e., consortia presentations and pre-read documents), including detailed project schedules. Without these detailed plans, TAC members find it difficult to understand the nuts and bolts of how the consortia operate and whether they are on track. Thus, the focus of the December 2021 TAC meeting will be on project management, with a secondary focus on psychometrics. The consortia are being advised to show the TAC the progress that has been made on comparability to date and describe plans to for the remainder of the pilot period. Specifically, the consortia have been asked to:

- describe the elements of the Comparability Guidelines for which they may already have sufficient evidence;
- describe the status of elements of the Comparability Guidelines for which they do not yet have sufficient evidence; and
- describe the plan, including the process and timeline, to develop sufficient evidence for the remaining elements of the Comparability Guidelines.

Technical questions will likely be raised as the consortia present documentation and describe future analyses. Ideally, preparation for the TAC will help the consortia refine their timelines and better understand the requirements that must be met in the next few years. Clear plans and timelines for establishing comparability will help assuage TAC concerns around progress. The quality of such plans will also signal to GaDOE whether additional technical assistance is likely to be needed during 2022 such that documentation can be appropriately evaluated by WestEd and the TAC.

SUMMARY

Throughout the first two years of the IAPP, both consortia were forced to pivot in response to the COVID-19 pandemic. Given that the assessment systems are locally supported, district needs were prioritized above meeting original project timelines. Thus, the consortia focused more on providing professional development to participating districts and keeping stakeholders engaged in the pilot than in field test completion. Although the work of building and scaling the assessment systems is now behind schedule, the 2021-22 school year might provide the data needed for the consortia to make more progress. It's possible that the consortia will be able to make up for lost time and get approval for use in place of Georgia Milestones by Year 5 of the IADA pilot period. However, it's quite likely that at least one of the consortia will need until Year 6. Additionally, only ELA and mathematics would be ready by Year 5 or 6—the implementation schedule for the other subject areas has been staggered such that it would likely be Year 7 or later before comparability evidence could be reviewed and the assessments could be approved for use instead of Georgia Milestones. Though the current IADA period ends after Year 5, the federal government has indicated that two-year extensions could be provided upon state request.

Delayed IADA timelines are not unique to Georgia. Many of the other states have faced similar setbacks due to the pandemic. Innovative assessment pilots in general have also taken longer than expected given the sheer complexity of running multiple assessment programs concurrently in a state and evaluating the outcomes of new assessment models. Even within the state of Georgia, updates to the content standards for math and ELA have been delayed a year. Nevertheless, stakeholders are interested in the continued viability of the IAPP in Georgia and are closely monitoring consortia progress toward operational administration.

Year 2 of the IAPP reflected many of the same challenges described in the Year 1 report:

- **Delays due to COVID-19's impact on the educational system.** Data was not available for Georgia Milestones in 2020, which delayed Year 1 of IAPP implementation. Although some data were available for Georgia Milestones in 2021, participation rates were much lower than normal and opportunity to learn was impacted by the ongoing pandemic. Thus, the consortia were not able to gather the data and conduct many of the analyses they had envisioned during the first two years of the pilot. The delays from Years 1 and 2 of the pilot will have a lasting impact on future years. If Year 3 participation rates on Georgia Milestones and the innovative assessments are reasonable, there is some hope that the consortia will be able to move forward, make up for some lost time, and successfully launch their innovative assessments in Georgia.
- **Resource constraints in terms of federal and state funding.** The consortia were not provided with funds to build and scale their assessment systems in Year 1, nor was GaDOE provided funding to oversee the project and review comparability documentation. In Year 2, Georgia allocated $250,000 to each consortium. Nevertheless, half a million dollars is nowhere near the amount of money spent on state summative assessment programs, so the consortia

must rely on funds from districts, philanthropies, and internal vendor resources. Furthermore, the two consortia are not equally funded or staffed. These challenges are unlikely to be resolved in future years of the pilot.

- **Inevitable challenges around the competitive design of the pilot.** Passionate local supporters of each assessment have invested significant time and energy into these projects. It's unclear how a single approved assessment at the end of the pilot will be accepted statewide. In the meantime, a firewall between the two consortia prevents sharing of ideas and lessons learned. This challenge has not changed from Year 1 to Year 2 of the pilot and will only become more pronounced as the two consortia scale and continue to invest in the process over time.

Additional challenges in Year 2 included:

- **Low usage of technical assistance.** Available technical assistance hours were cut back dramatically in Year 2 of the pilot, but the consortia made limited use of the available hours. With the prospect of data in Spring 2022, the consortia may have more detailed technical questions around analysis plans and results and may need to request additional technical assistance.
- **Challenges around TAC preparations and consortium project management.** The technical assistance provided by the TAC is most useful when the TAC has had time to review materials ahead of time and think through advice. GMAP submitted materials ahead of time as requested, but Putnam often struggled to get materials submitted prior to the TAC meetings. Both consortia would benefit from including information in TAC materials around what feedback they heard from the TAC previously, what they've done to address the feedback, and rationales for when they decided not to implement feedback. The TAC also expressed concern about whether Putnam had a workable project schedule and process for tracking all aspects of what will become a complex enterprise as the consortium moves toward operational administration in multiple grades and subjects.
- **Progress and decision-making.** Progress has been slow and many decisions that needed to be made at the outset of the pilot are still outstanding decisions at the end of Year 2. Delays are understandable given the context of the last two years. However, additional progress on analysis plans and development of potential solutions for the various outstanding decision points (e.g., what to do for a CCRPI literacy measure) might have been possible.
- **Lack of experience with accountability assessments.** GMAP's vendor, NWEA, is not a newcomer to large-scale assessment. Their interim assessment products are used nationwide. What is new for NWEA is creating a customized solution for a specific state that will meet state and federal accountability requirements. Putnam's vendor, Navvy, has much more limited assessment experience as a fairly new company which developed a Georgia-specific formative assessment. Thus, the Putnam team has a learning curve involved with both large-scale assessment and the accountability systems into which the assessment results must fit. As newcomers to the statewide summative assessment space, the consortia often have questions about the constraints of the existing accountability system. They have

benefitted from access to technical assistance provided by the TAC and WestEd, who have helped them ask questions that were not immediately obvious and point out aspects of the process have been underestimated.

- **Justifying differences between the innovative assessments and Georgia Milestones.** The TAC has noted on many occasions that differences between the innovative assessments and Georgia Milestones are potential sources of non-comparability. Thus, the TAC's advice is often to use the same procedures that have been used with Georgia Milestones previously. For example, the process used to establish alignment of Georgia Milestones to Georgia's content standards is quite likely a good process to use with the innovative assessments. Of course, if all aspects of the innovative assessments matched Georgia Milestones, then there would be no innovation. Nevertheless, differences between the two assessments must be justified based on theories of action and theories of learning. For example, testing at the end of the year makes implicit assumptions about measuring the retention of learning, while through-course assessment measures learning as it happens, but may not reflect the total amount of knowledge a student retains at the end of the year. These theories should be tested with empirical data as it becomes available. The unintended consequences of end-of-year assessments are in large part due to the high-stakes decisions made based on test scores. Once through-year assessments are used for the same high-stakes decisions, the same unintended consequences might result.

Innovation is not expected to be easy, and when high-stakes decisions and multiple stakeholder groups are involved, innovation is also not likely to occur fast. Thus, it will be important to track whether the required investment of time and resources results in an improvement in the education of Georgia's students.

Lessons Learned and Next Steps

Year 2 of implementation of the IAPP was not necessarily smooth, but progress is being made. Innovation rarely happens overnight; rather, it takes many years to build new systems. Although comparability is the ultimate criterion for IADA, the real test for the consortia will be the outcomes for students, teachers, and schools once comparability is established.

The past year has highlighted areas where additional planning is needed and where important decisions remain. In Year 3, more data will be available to inform some of these decisions. Moving forward, the consortia should leverage the expertise of the TAC and WestEd's technical assistance to make additional progress on the following technical components of their assessments:

- Finalizing the process for calculating the student scores that will feed into the accountability system
- Finalizing plans for selecting an external alignment evaluator and carrying out alignment studies

- Finalizing analysis plans for Spring 2022 data (and future data collections)
- Identifying potential CCRPI growth and literacy measures and developing plans for choosing a method from among various options
- Creating business rules for defining participation (e.g., how many testing events or questions must a student complete?) as well as establishing procedures to handle cases where students move into the district or state mid-year
- Refining theories of action and plans for evaluating the claims the consortia want to make about their assessments (e.g., does a through-year model change instructional practice?)
- Refining the plans and the schedule for submitting documentation required in the Comparability Guidelines
- Continuing item development and item review for new grades and subjects (i.e., science and social studies)

Building on the Comparability Guidelines which were developed in Year 2, WestEd and GaDOE will develop a process for the collection and review of comparability evidence so that the multi-step review process can be implemented efficiently beginning in Year 3 and continuing into Years 4 and 5 and the state can realize the goals of the IADA process.

APPENDICES

Appendix 1: Technical Advisory Committee Meeting Summaries for Putnam County Consortium and Georgia MAP Assessment Partnership, December 2020 and July 2021

Appendix 2: Georgia Innovative Assessment Pilot Program Assurances

Appendix 3: Georgia Innovative Assessment Pilot Program Comparability Guidelines

Appendix 1

Technical Advisory Committee Meeting Summaries for Putnam County Consortium and Georgia MAP Assessment Partnership, December 2020 and July 2021

Georgia Innovative Assessment Pilot Program

DECEMBER 2020 TECHNICAL ASSISTANCE COMMITTEE MEETING

Georgia MAP Assessment Partnership

Markie McNeilly
Matthew Gaertner
Assessment Research & Innovation @WestEd | csaa.wested.org

GEORGIA INNOVATIVE ASSESSMENT PILOT PROGRAM

DECEMBER 2020 TECHNICAL ASSISTANCE COMMITTEE REPORT FOR THE GEORGIA MAP ASSESSMENT PARTNERSHIP

INTRODUCTION

The Georgia Innovative Assessment Pilot Program (IAPP) Technical Advisory Committee (TAC) meeting was convened on December 15, 2020. The meeting was held virtually via Zoom video conferencing. Attendees included members of the TAC, the Georgia MAP Assessment Partnership (GMAP), NWEA, 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 CONSORTIUM ASSESSMENT SYSTEM

DESCRIPTION

The GMAP Partnership and NWEA provided an update on the consortium's assessment system. The COVID-19 pandemic shifted the timeline for planned activities. The consortium shared details on the continued impact of the COVID-19 pandemic. Most notably, the decision was made not to field test in Spring 2021, as previously planned. The overall timeline for producing an operational test and for establishing comparability has been shifted out by at least a year. They also shared updates on the consortium's membership as well as status updates on content development activities, psychometric activities, and the development of student score reports.

TAC DISCUSSION AND RECOMMENDATIONS

During the presentation, the GMAP Partnership shared that two new districts were approved by the consortium to join their membership — Chattahoochee and Calhoun. Both of these districts are in the southeastern area of the state, which has not been represented in their membership until now.

An update was given on content development activities. ELA and math items are in development, with the first field test planned for spring 2022. They are working on the range PLDs as far as they can at this point in their process. They worked with their content advisory boards (composed of educators from across the state) to review the new assessment items. Item content and bias reviews took place over the summer. Science development — the first draft of range ALDs and item specifications — are in development. Content development activities will continue, with additional review committees planned for next summer.

Within the field test plan, references to open-ended questions in the writing domain have been removed. Items requiring hand-scoring have been deferred, and the consortium will revisit their

inclusion once the test becomes operational. Instead, technology-enhanced items will be included to measure writing. Technology-enhanced items are multiple-part items that measure aspects of the writing process, without requiring students to actually write. These item types have been used for a few years now. One of these item types includes highlighting text within a passage. The TAC would like to see what these items look like at a future meeting.

Psychometric activities have also progressed. NWEA has been working on how technology and processes will need to be set up in order maximize valid and reliable results. They have been conducting item calibration studies and optimizing code. A range achievement level descriptor (RALD) utility study is underway, but it has been difficult to progress without being able to get into classrooms. NWEA has also been working through vetting the spring 2022 field test plan. Through-year Computer Adaptive Test (CAT) simulation studies have been conducted and will continue over the next year.

NWEA provided an update on the development of a family score report. A prototype was reviewed by GMAP districts over the summer. A usability study was conducted with parents/guardians and teachers in the fall. Score report prototypes will continue to iterate, incorporating information and feedback from stakeholders (teachers, students, families). Participation in the score report activities over the summer was limited to three of the member districts due to the pandemic. As students return to the classroom, engagement is slowly increasing. The TAC would like to see what the score reports look like at a future meeting.

Field Test Plan for Spring 2022

Description

NWEA shared an update on the field test plan for the ELA and math assessments, now projected to take place in spring 2022. The basic field test design, content design, and timeline were presented.

Students will take MAP with field test items included. The test will be longer than a typical testing event because MAP results still need to be produced, including a RIT score which many schools utilize for student classification. Reliable summative scores will also need to be produced. This will happen after the field test data have been calibrated. Further, a comparability study is planned for summer 2022. Sufficient field test items must be administered in order to have an operational test in spring 2023. The TAC suggested that NWEA develop and evaluate success criteria for the field test when finalizing their plans.

TAC Discussion and Recommendations

Learning loss due to COVID-19 was discussed. It is unknown how student performance on the assessment will be impacted by learning loss from the 2020-2021 school year. NWEA plans to evaluate the stability of the scale each year and if necessary, recalibrate and rescale.

MAP Growth will be administered in fall and winter of 2021-2022 within the typical timelines and the usual technology platform. In spring 2022, students will take the regular MAP Growth and adaptive MAP Growth tests on a new platform. The TAC recommended trying to get a measure of student

motivation (such as item latency and completion rates). Additionally, they suggested getting feedback from teachers and students about their experience and how much effort they exerted on field test items.

Sample items will be made available ahead of the field test, since field test items will look different from the MAP Growth items students are used to seeing. The TAC supports this approach, and also recommended including sample items in the beginning of the test. Including sample items in the beginning of the test will ensure that all students have an opportunity to practice interacting with the technology-enhanced items.

The TAC had some concerns over the number of items that are included for field testing. NWEA explained their field-testing approach including limitations on the number of participating students and the number of items needed to support an operational CAT item pool. The TAC recommended reducing the number of items students are given in the field test as much as possible, be it through increased recruitment or otherwise. The TAC also suggested finding alternate solutions to embedding the field test items on the test. One suggestion included embedding or partially embedding field test items within the MAP Growth test. In this way, it is less obvious to students that these are items that do not count toward their score. Another recommendation was to provide different forms to students, so that on some forms the field test questions would appear first and on other forms the field test questions would appear after the MAP Growth test.

Suggestions from the TAC also included altering the design of the field test. For example, NWEA could consider eliminating the GMAP individual-level summative score during the field test in order to reduce the number of items administered to each student. Decision consistency across Georgia Milestones and GMAP could be projected based on aggregate level data.

During NWEA's high-level overview of the field test design, NWEA and the TAC discussed the placement of item blocks within a form. The TAC recommended constraining passages to a specific location in the operational delivery. Another option is to constrain the number of passages and fix them within two slots on the test form. There may be value in varying the location of the passage blocks because the item positions will vary on the adaptive test.

NWEA asked for the TAC's advice on how to place ELA and reading items into an existing reading scale if they use a fixed-person parameter calibration. The TAC recommended that NWEA verify the approach and that the theta scores that are generated are either equivalent or close enough to be considered comparable. The TAC suggested that it may be helpful to look at the stability of the theta estimates for a 30-item MAP Growth test versus a 40-item MAP Growth test. The TAC said that there might be a dimensionality issue; however, there are a number of other assessments that have used this same approach (e.g., ELPA21, CPA exam).

NWEA asked for the TAC's recommendation on how to approach the reading scale if the correlation doesn't support a claim that they are equivalent or around the same scale. While the TAC acknowledged that both a reading RIT score and a GMAP ELA score could be provided. The TAC

encouraged NWEA to consider other models moving forward, especially if open-ended writing items are eventually added to the mix. The TAC had some concerns about using TEIs in place of writing prompts, noting that there may be unintended consequences of using different measurement approaches even when the scores are highly correlated.

When reviewing the field test timeline, the TAC recommended to prioritize tasks based on goals, identifying activities that could be scaled back or eliminated so that the project can be maintained despite the multitude of external factors in play this year. At future meetings the TAC would like an update on the field test plan as well as an opportunity to view the MAP Growth reports and any prototypes of the summative GMAP score reports, if available.

Technical Criteria for Evaluating Field Test Items

Description

The GMAP Partnership and NWEA presented the criteria that they plan to use to analyze field test data that has been collected. The presentation included information on calibration procedures, vertical scaling, and the data review process. They requested the TAC's feedback on the criteria and process that they have developed.

TAC Discussion and Recommendations

NWEA asked if the TAC had any recommendations that they should consider for item flagging criteria, including fatigue and motivation effects on item performance. The TAC noted that item difficulty can be affected by item position and the context effects of having different surrounding items. If possible, vary the position of items across forms and evaluate the impact on item difficulty estimates. If the item difficulty looks extremely different, then the item should be considered for removal from the item pool. The TAC also recommended to incorporate Steve Wise's research on measuring student effort and engagement.

Comparability Evidence and Timeline

Description

GMAP and NWEA presented information on comparability. They are planning on doing the bulk of the empirical data analysis for comparability in the summer of 2022. There are some activities, such as establishing content comparability and alignment evidence, that they will be able to complete ahead of time. Their goal is to establish score comparability between GMAP Summative and Milestones, as well as between GMAP Summative and MAP Growth. Comparability between GMAP and MAP Growth is desired by the GMAP Partnership school districts, as they can continue to have the ability to use all of the RIT scores for the same purposes they have used them in the past.

TAC Discussion and Recommendations

GMAP reporting will provide a growth measure and a summative measure. The current plan is to use the MAP RIT scale as the measure of growth. NWEA is also looking at comparability between

Milestones and GMAP at the classification level — where students will be classified into comparable achievement levels. This is in alignment with what has been discussed at previous TAC meetings.

The TAC noted that the consortium should be able to get a good projection for comparability as long as they have a representative sample. The GMAP Partnership should also be prepared to show that they've done an alignment study that shows the content is comparable, and that they have looked at it empirically.

NWEA noted that they have already conducted a linking study between MAP Growth and several state assessments, including Georgia Milestones. However, MAP Growth is not well aligned with the Georgia content standards and assesses off-grade level content. GMAP is specifically aligned to the Georgia content standards, measuring on-grade level content only, so a comparability analysis between GMAP and Georgia Milestones is needed.

The blueprints between GMAP and Milestones are very similar in terms of proportions of items and reporting categories. There are differences because GMAP is an adaptive test. NWEA described a plan to create a binary classifier to find the cut scores on GMAP that correspond to the cut scores on Milestones so that the classification agreement is maximized. However, the use of logistic regression would create an asymmetric relationship between the two cut scores. A symmetric function, for example equipercentile linking, would be preferable.

NWEA discussed the design for data collection. There will be a naturally occurring counterbalanced design for the order in which students will take Milestones and GMAP because districts are already approaching this differently. Some students will take Milestones first and others will take GMAP first. The TAC noted that if the sample is not equally representative of the population, NWEA may want to utilize weights to better approximate the population in the counterbalanced design.

The TAC recommends replicating the comparability study as the number of participating districts grows and becomes more and more similar to the statewide student population.

Plan and Timeline for Releasing Items

Description

NWEA presented plans and timeline for releasing items. An item sampler/GMAP tutorial is being created for students to be able to get familiar with where tools are located, how to interact with items, and how to advance through the assessment. Additionally, previously tested items will be released to provide additional examples of the content that is on the test for students, teachers, etc.

TAC Discussion and Recommendations

NWEA is estimating that they will release 10 items per year, per content area, per grade. In the future, once the bank is larger, they may be able to increase the number of items in order to get a better distribution of the content. Scoring information will also be provided so that students can check their answers. Data will be shared for released items, such as standard alignment and justification for why

they were chosen. The TAC suggested that it would be helpful for practitioners to have more information about the released items, such as their difficulty level and the difference in performance across proficiency levels. The TAC also recommended that there be at least two items per technology-enhanced item type in the sampler so that students have multiple opportunities to practice using each item type.

Next Steps

TAC Requests

At the conclusion of the TAC meeting, the TAC requested that the following be addressed in future meetings:

- An update on the range ALDs
- A theory of action, including discussion on the assessment's intended impact on teaching and learning
- An update on alignment studies and their results
- Additional information on score reporting and its links to professional learning for educators

During the TAC Debrief between the TAC, GaDOE, and WestEd, the TAC requested the following from each of the consortium:

- Provide a summary of key takeaways and action items from the TAC meeting to the TAC.
- During the summer 2021 TAC meeting, discuss the outcomes of the recommendations provided by the TAC in this meeting. Provide information or justification if recommendations were not taken.

Georgia Innovative Assessment Pilot Program

DECEMBER 2020 TECHNICAL ASSISTANCE COMMITTEE MEETING

Putnam County Consortium

Markie McNeilly
Matthew Gaertner
Assessment Research & Innovation @WestEd | csaa.wested.org
January 2021

GEORGIA INNOVATIVE ASSESSMENT PILOT PROGRAM

DECEMBER 2020 TECHNICAL ASSISTANCE COMMITTEE REPORT FOR PUTNAM COUNTY CONSORTIUM

INTRODUCTION

The Georgia Innovative Assessment Pilot Program (IAPP) Technical Advisory Committee (TAC) meeting was convened on December 14, 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 CONSORTIUM ASSESSMENT SYSTEM

DESCRIPTION

The Putnam Consortium and Navvy Education provided an update on the consortium's activities and development of the Navvy assessment system. The consortium shared details on the continued impact of the COVID-19 pandemic. Two staff members from Scintilla Charter Academy, Amanda Dean, Assistant Dean, and Brooke Night, an Instructional Guide, joined the meeting to share their experiences using Navvy in their school. They shared what the system looks like and what feedback it provides as they track their students' progress throughout the year.

TAC DISCUSSION AND RECOMMENDATIONS

The Putnam Consortium shared information on challenges schools faced returning for a new year amidst the COVID-19 pandemic. Schools navigated providing options to families for in-person, online, and hybrid learning, particularly for low-income and rural families for whom connectivity has been a challenge. Schools are able to administer Navvy, but they have chosen to do so in varying degrees. For example, some schools have only chosen to administer the assessment for a selection of standards, while others are committed to administering the assessment for every standard.

The schedule for conducting comparability analyses was similarly delayed. Putnam now plans to use student results from the 2020-2021 school year and conduct a comparability analysis with a representative sample of students (assuming Georgia Milestones is administered). This activity was originally planned to take place in the 2019-2020 school year but was postponed due the pandemic. Had there been no delays, Putnam County would have run a check of the comparability during the 2020-2021 school year.

The TAC discussed the use of Navvy data from the 2020-2021 school year. Given the disruption to instruction and new and differing opportunities for learning, the data may show that students experienced some learning loss. The TAC suggested that the data can still be used as a valid measure of achievement and can be used to see how students and teachers are performing under current conditions. The data probably will not support cause-and-effect claims, though, because some students are not receiving the same opportunities as others (e.g., some students are still in completely online classroom environments). In other words, datasets will need to be contextualized within the circumstances of the districts they are coming from.

During this discussion, educators from Scintilla Charter Academy provided insight into their experiences using interim assessment systems and shared the value they perceive in using the Navvy assessment system. Putnam shared that parents are able to log in to the system as their student to see their scores and progress. The TAC recommended that the Putnam Consortium establish a method to ensure students understand what each standard is asking of them. Suggestions included conducting a small cognitive lab, including a released item with each standard, and rewriting the standards to create an unofficial copy without educational jargon.

Review of Communication Materials

Description

During the June 2020 TAC meeting, the Putnam Consortium received feedback on strategies for scaling up the assessment system and recommendations on communication materials. The Putnam Consortium presented their progress on the communication materials during this session. The TAC provided further feedback on the presentation of the materials and strategies for communicating with stakeholders about the assessment system.

TAC Discussion and Recommendations

In response to TAC feedback, Navvy produced a checklist to share with various stakeholders that compares the Navvy assessment to other interim assessments that districts may be utilizing in Georgia. This tool serves as a method to explain how Navvy differs from the other products. The TAC recommended that Navvy share the checklist with the developers of the assessments on the checklist to ensure their assessments are accurately represented. The TAC also suggested organizing the descriptors by audience (some descriptors will be more relevant to parents, some to administrators, and so on). Additionally, the TAC recommended emphasizing the reports that Navvy produces when marketing the assessment to the field; stakeholders will likely perceive the information those reports provide as valuable.

Evaluation of Navvy Assessment System Effectiveness Plan

Description

The Putnam Consortium has designed a study to help understand the impact the Navvy assessment has on teaching and learning based on feedback received from the TAC at the last convening. The design matches schools that are administering Navvy with schools that are not administering it based

on a number of variables, such as demographics and past student performance. They plan to compare results from the Milestones summative assessment between the matched schools. This is not a requirement of the Innovative Assessment Demonstration Authority; however, Putnam argued that this study will help ensure the assessment system is working and will provide valuable information to stakeholders considering participation in Georgia's Innovative Assessment Pilot Program with the Putnam Consortium. With the understanding that this year's data collection and use may look different than in upcoming years, the Putnam Consortium requested feedback from the TAC on the design of this study.

TAC Discussion and Recommendations

The TAC recommended considering the different learning models that are taking place in each of the schools when matching schools and analyzing data. They also suggested amplifying the theory of action for the study by considering three components that are needed to be successful in order to support their claim: assessment results, teacher capacity to utilize results, and differential approaches to instruction. They suggested that because data may not be generalizable for this year, the Putnam Consortium may want to focus on a narrow case study with a small group of teachers who have been using the results to personalize instruction.

The Putnam Consortium included a brief review of literature conducted to help inform the study, noting that they were not able to find much research on how assessment systems help students learn. They noted that there is a body of literature on data-based decision-making and how interim and benchmark assessment can predict summative assessment results. The TAC recommended the Putnam Consortium review research reports published by Smarter Balanced, Regional Education Lab reports on formative assessment, and works by Joan Herman and Suzanne Lane.

Comparability Discussion

Description

The topic of comparability surfaced throughout the meeting. The Putnam Consortium understands that the requirement is to roll up the data from Navvy to provide an annual summative determination for each student, which needs to be comparable to the achievement level the student receives on the Milestones assessment (this is the current statistical comparability threshold, one of many pieces of evidence required before an assessment can be administered in lieu of Milestones; the TAC will take up this topic during the spring/summer 2021 meeting). There are two approaches they are considering for establishing the summative determination: either to maintain the multivariate profile of standards competency or to consolidate the multivariate profile into a single numerical result. The TAC's feedback on which approach to utilize was requested.

TAC Discussion and Recommendations

The TAC recommended that the Putnam Consortium try both approaches for obtaining comparability evidence. Consolidating the multivariate profile into a single numerical result may be fruitful because the scores can more easily be mapped back to the Milestones test specifications. They reiterated that

the test needs to be comparable at the performance level, and not at a finer grain of detail, because the tests are different. They also indicated that validity evidence is also needed when establishing comparability.

In addition to score comparability, the system must also have comparable supports to the statewide assessment system. For example, the system must have adequate test security, appropriate and reasonable accommodations for students, and alternate methods for assessing students with significant cognitive disabilities. These elements were described by each consortia in their initial application for the innovative assessment program.

Science Partners

Description

The Navvy assessment has been built out for ELA and mathematics subject areas. Development of the science assessments has not yet begun, and Navvy is looking for partners to help in this effort. Navvy asked the TAC if they had any recommendations for groups that are currently working in science assessment that would be beneficial to speak to.

TAC Discussion and Recommendations

The TAC provided the names of test development companies that Navvy Education could consider reaching out to. Navvy and the Putnam Consortium encouraged the TAC to reach out if they think of any other groups after the meeting had concluded.

Next Steps

TAC Requests

At the conclusion of the TAC meeting, the TAC requested the following be addressed in future meetings:

- Present Navvy's theory of change and how it relates to the challenges faced due to the pandemic.
- Address where activities lie on the continuum of development and how the pandemic has shifted these activities. Share what had to be postponed and what will need to be redone.
- Provide TAC the meeting slides and any supplementary materials at least one week before the TAC meeting takes place.

During the TAC debrief between the TAC, GaDOE, and WestEd, the TAC requested the following from each of the consortia:

- Provide a summary of key takeaways and action items from the TAC meeting to the TAC.
- During the summer 2021 TAC meeting, discuss the outcomes of the recommendations provided by the TAC in this meeting. Provide information or justification if recommendations were not taken.

Georgia Innovative Assessment Pilot Program

JULY 2021 TECHNICAL ASSISTANCE COMMITTEE MEETING

Georgia MAP Assessment Partnership

Mariann Lemke
Sonya Powers
Assessment Research & Innovation @WestEd | csaa.wested.org

GEORGIA INNOVATIVE ASSESSMENT PILOT PROGRAM

JULY 2021 TECHNICAL ASSISTANCE COMMITTEE REPORT FOR THE GEORGIA MAP ASSESSMENT PARTNERSHIP

INTRODUCTION

The Georgia Innovative Assessment Pilot Program (IAPP) Technical Advisory Committee (TAC) met on July 7, 2021, via Zoom video conferencing. Attendees included members of the TAC, the Georgia MAP Assessment Partnership (GMAP), NWEA, the Georgia Department of Education (GaDOE), and WestEd. EdMetric also attended for part of the meeting to describe their alignment work on behalf of GMAP. The agenda included two main topics:

- a review of comparability requirements and associated discussion of their specific application to the GMAP assessments; and
- an update on GMAP's implementation.

This report provides an overview of each topic and a description of the resulting key takeaways and action items from the meeting.

COMPARABILITY REQUIREMENTS CHECKLIST

To begin the meeting, WestEd staff provided an overview of the comparability evidence that each consortium will be required to provide to the state. Examples of relevant evidence are described in a template that will be provided to GMAP. Evidence is required in several main categories, as described in the following sections.

Alignment and Comparability

Consortium assessments must demonstrate that:

- assessments and items are aligned to the Georgia standards,
- assessments match the depth and breadth of the Georgia standards,
- students can be classified into at least four achievement levels representing the same knowledge and skills that current Milestones assessment achievement level descriptors (ALDs) provide,
- summative classifications of students are consistent across Milestones and innovative assessments (for all students, subgroups of students, content areas, and assessments),
- those who participate in the innovative assessment are representative of the state in terms of demographic composition and achievement, and
- there is a plan for conducting annual comparability analyses between the innovative assessment and Georgia Milestones throughout the remainder of the IADA period.

To meet these criteria, the consortium should present an independent alignment study including information similar to that provided in previous Milestones reports. Four types of alignment should be included: balance of complexity, depth and range of knowledge, and categorical concurrence. Note that conducting an alignment study of all items is not necessary (though every grade level should be included). A sampling approach that provides strong evidence that the items and tests that students actually encountered on a consortium assessment are aligned (for example, by selecting a sample of students across proficiency levels and checking alignment for those students' tests) can suffice. Note also that the state is updating its standards. New math standards will become operational in 2023–24 and ELA in 2024–25, so new evidence of alignment will be needed after the new standards become operational.

The consortium must also demonstrate that it has achievement levels that correspond to the current Milestones ALDs. Direct adoption of Georgia's ALDs can satisfy this criterion, though other ALDs may be used with evidence of their alignment to the existing ALDs. The consortium must show evidence that students at each of the Milestones ALD levels have the skills and knowledge described in those ALDs. For example, if the Milestones ALD describes proficiency as being able to use place-value relationships to round numbers, the consortium should demonstrate that students placed into that performance level on the innovative assessment also demonstrate those skills.

The consortium must also provide a report on how classification into its achievement levels compares to classifications on the Milestones assessment. Only on-grade-level items should be used to classify students into performance levels. It is possible that new tests may provide different results for good reasons, based on the design of the assessment or the approach to scoring; the consortium should be prepared to fully explain and justify why differences may occur. The consortium should be sure to describe not just how many students are at each level but the degree to which students are consistently classified by the two assessments. Because end-of-course assessments contribute 20% to course grades, the consortium should also provide evidence of its approach to using its scores for grades and the comparability of those grades to the grade conversion score (GCS) method used with the Milestones assessments.

Consortium documentation should also include descriptive analyses of its participating populations of students, compared to the state, with description of weighting methods or other mechanisms for generalizing sample results to the state, as relevant. All state-reported subgroups of students should be included, as well as a description of groups based on achievement.

Beyond initial comparability analyses based on students taking both the consortium assessments and the Milestones tests, the consortium must provide a plan to conduct annual comparability analyses for the remainder of the IADA period. This plan need not include testing of all students, but, rather, should include a sample of grade bands (or grade bands/students), so that each grade band includes an innovative assessment and the state assessment (see IADA final regulations, pp. 28–29).

Technical Quality

The consortium must also provide evidence of the technical quality of its assessments, demonstrating:

- work with experts to ensure quality,
- reliability and validity of the assessments,

- how the assessment provides information across the full performance continuum for students,
- availability of individual and aggregate reports and the timeliness and interpretability of these reports for stakeholders,
- how principles of universal design for learning were incorporated into the assessment design, and
- a plan to maintain the item bank and the integrity of the score scale over time.

To meet these criteria, the consortium should provide background information (e.g., names, CVs) of TAC members and agendas of meetings aimed at discussing technical quality of the assessments.

The consortium should also present evidence of validity that matches the categories in the *Standards for Educational and Psychological Testing.* Not all evidence (e.g., consequential validity) may be available immediately, but the consortium should describe its plan to gather this information over time. Consideration of what validity evidence can be provided without testing, what can be gathered during piloting, and what must be gathered once an innovative assessment is fully operational may be useful.

The consortium must provide reliability evidence for the summative scores, subscores, and achievement levels generated from the innovative assessment, consistent with national standards and the Georgia Milestones. For example, evidence might include test-subtest reliability (again, including only on-grade-level items). Decision consistency and accuracy values should be similar to those reported for Georgia Milestones.

Data showing the distribution of scores, to demonstrate how the assessment provides information across the performance continuum, should also be presented. These data could include analyses of test information functions or other analytics, or other types of information such as cognitive lab data and test blueprints indicating depth-of-knowledge ranges.

The consortium should provide examples of its student and aggregate-level reports (such as classroom, school, consortium, and even state-level reports). These reports should be accompanied by evidence that stakeholders can use these reports to make valid interpretations about student performance, such as data drawn from focus groups of a variety of stakeholders representing report consumers, data from A/B tests, or other data.

Innovative assessment reporting timelines must describe when and how stakeholders receive results of the assessment, demonstrating that these results are provided in a timely manner. Final results for accountability must be provided at least in the same timeframe in which the current Georgia Milestones assessment final results are available.

The consortium should also provide a description of how its assessments incorporated principles of universal design for learning in test development, as well as how scales and item banks will be maintained over time (e.g., how parameter drift will be managed).

Accessibility and Accommodations

All students who currently participate in Georgia Milestones must be able to participate in the innovative assessment in order to use the innovative assessment in lieu of Georgia Milestones, including students with disabilities and English learners (except students with the most severe cognitive disabilities, who may participate in an alternate assessment).

A crosswalk of accessibility and accommodation features available on Georgia Milestones and available on the innovative assessment should be provided such that it is possible to see, at a glance, whether all of the accessibility and accommodation features will be available, and, if not, how students will be validly assessed using an alternative accessibility mechanism. Any differences in the ways that accessibility or accommodation features work in the innovative assessment, compared to Georgia Milestones, should be indicated.

Accessibility features and accommodations must allow students to participate in alignment with their IEPs or English learning plans and comply with relevant federal laws such as the Individuals with Disabilities Education Act (IDEA). The consortium should provide a participation report that shows that all students are participating as required.

The consortium need not have all accommodations available in order for the innovative assessment to be approved for use in lieu of the Georgia Milestones, but must have a specific and feasible plan to provide all needed accommodations when assessments are administered. For example, the consortium need not have Braille forms ready at the time that evidence of comparability is being reviewed, but must have a well-described plan to produce Braille forms prior to administration, that demonstrates the vendor's capacity to produce them (historical evidence of how they have been produced in the manner described).

Test Administration and Security

The consortium must demonstrate that it has plans in place to ensure standardized administrations, such as training and manuals, and processes to prevent and/or document testing irregularities and protect test security and student data. In addition, the Georgia Office of State Assessment will monitor consortium test administrations, and monitoring reports should be included in evidence for this criterion. Other evidence would be sample irregularity reports, results of analytical analyses aimed at discovering cheating, auditing procedures, and procedures to handle irregularities or test security violations.

The consortium should keep in mind that standardization processes are intended to promote the validity and comparability of the scores, but the consortium need not compromise features of the assessments that make them innovative. As an example, using many different types of accommodations reduces the standardization of administration, but is necessary to ensure validity of the scores.

Stakeholder Engagement

The consortium should provide evidence that assessments were developed 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. Evidence might include letters of support or agendas from meetings where assessments were discussed, along with participant lists.

The consortium should also document how it has worked with schools and districts to interpret results and communicate with stakeholders such as parents, students, and community members (i.e., how the consortium has worked to develop assessment literacy). Evidence might include training agendas and presentations, meeting agendas, assessment guides, score interpretation guides, data on stakeholder participation in training for test administration or score interpretation, or stakeholder survey or focus group data.

Accountability

Georgia's accountability requirements must be met with use of any innovative assessment. In addition to the need to provide a summative score, these requirements also include providing measures for the College and Career Ready Performance Index (CCRPI).

The consortium should demonstrate that it uniquely identifies students within and across years so that students' assessment data, schools, districts, demographic information, etc., can be used for accountability purposes. Data layouts and timelines should be provided. Evidence must also be provided that the percentage of students assessed is at least as high as the percentages observed on Milestones prior to the start of the innovative pilots, overall, as well as for all federally required student demographic subgroups.

The consortium must describe how it will produce a single summative score. If there is more than one administration during the academic year (e.g., a through-year model), the consortium should specify which administrations contribute to the summative score and how scores are combined. This description should provide a clear rationale for the calculation of the summative score.

As noted, the consortium must also show how its assessment data can be used for a variety of CCRPI purposes, including providing measures for the Content Mastery and Closing Gaps components of the index, growth measures for the Progress component, and literacy measures for the Readiness component. These measures do not need to be strictly comparable to, or use the same methods as, the Georgia Milestones, but evidence must be provided that justifies the proposed approach.

Conflict of Interest

The consortium must provide assurances that there are no conflicts of interest (financial or otherwise) for parties participating in the pilot program, and that all local procurement rules are being followed. No new evidence is needed unless there have been changes since initial assurances were made at the award of the innovative assessment grants.

TAC Discussion and Recommendations

The TAC noted several aspects of the comparability requirements that the consortium will need to carefully consider, including the following:

Content Alignment

The TAC would like to see a traditional content alignment study where the GMAP items are aligned to Georgia content standards. NWEA described its range ALDs as an approach to keeping GMAP and Milestones comparable, but the TAC was concerned that differences between GMAP and Milestones ALDs might cause misalignments. The consortium would need to explain why the GMAP range ALDs are different than those used for Milestones. The TAC also reiterated that comparability is at the achievement level rather than at the scale-score level. The previous MAP alignment study is not sufficient because MAP was not created to be aligned to the GA content standards, but GMAP was developed to align to the GA content standards.

Reliability

GMAP asked about the reliability thresholds at the total test and subscore levels. The TAC would like information about how reliability and measurement error is calculated, and how statements about what students know and can do are justified, especially in terms of instructional recommendations. Milestones' overall reliability is around 0.9, so that should be the target for GMAP, but subscores will not have an official threshold.

Test Security

GMAP asked whether the administration security would need to be equally rigorous across all administrations if some of the administrations do not contribute to the summative score. The TAC mentioned that item exposure is a concern unless the item pool for summative scores is kept separate from item pools used for low-stakes administrations. All items that contribute to a student's summative score must be kept secure. Otherwise, having lower security for the interim assessments might be sensible.

Growth Measure and Score Comparability

GMAP asked whether its growth measure has to be the same as what is currently used by Milestones. GMAP can innovate and does not need to use student growth percentiles, but it should justify why a different method is used, and compare the results to Milestones to identify whether the results are different. The TAC noted that, ideally, student results would be the same regardless of which assessment they would take. If the metrics are not comparable, then which assessment students take will not be a matter of indifference. However, the purpose of IADA is to do something new, so changes that improves scores should not be eliminated. Any differences need to be explained, and if the differences are a reflection of something better, they are justified. Comparability is important because scores will be compared, and if there is a lack of comparability, it should be consistent with the theory of action.

"Banking" Scores and Score Interpretations with Ongoing Assessment

GMAP asked about the claims that one can make with a through-course model where the summative score is collected prior to the end of the school year. Is there a validity issue around what students have retained by the end of the year, versus the highest score the student attained across the school year? GMAP is still considering whether it might be possible to bank scores, but there is concern about validity and even comparability issues, compared to the Milestones model. GMAP has modified the through-year CAT design such that banking of scores would be possible. The blueprint for each assessment will be consistent across fall, winter, and spring. It is not designed to follow the scope and sequence in Georgia. The TAC indicated that this design would be more amenable to a score banking approach. To ignore the information gathered throughout the year does not make sense. Students who did poorly prior to the spring assessment should not begin at the same place as students who did well prior to the spring assessment. GMAP should capitalize on its adaptive technology. To meet accountability requirements, however, GMAP will need to represent the on-grade-level content. GMAP must clearly describe what a score is intended to mean. The assessment design does produce scores with different meanings and that will support different interpretations, but ultimately the consortium must be able to make the same claims that Milestones makes about students and scores.

Comparability Requirements Overall

The TAC recommends considering what is reported when providing validity evidence. Are the claims about what students know and can do substantiated?

The TAC recognizes that innovation may be difficult with the constraint of also meeting stringent comparability requirements. If it can be demonstrated that an assessment is of greater diagnostic value and instructional value, the TAC would take that into consideration when evaluating comparability evidence. However, the TAC also noted that the current comparability checklist is the bar to meet under current IADA requirements.

Update on Consortium Assessment System and Field Test Plans

During this part of the meeting, NWEA provided an update on work that GMAP has recently accomplished and work that is in progress, including information on recruiting and field test plans. Changes to the team were described, and new districts that have joined the consortium were named. Other updates related to the field test included GMAP's plans to:

- provide a reliable linked-RIT score;
- evaluate within-year and across-year growth;
- develop new reports rather than using MAP Growth reports (there is a new platform that will be used, requiring the move to the new reports);
- use assessments for determining eligibility for gifted programs;
- provide reliable GMAP summative scores with delayed scoring (late summer 2022), to be used in comparability;
- field test enough items in spring 2022 to create the operational through-year CAT with 50–60 items (more students able to participate);

- move forward with item-level CAT, rather than multi-stage adaptive;
- use theta estimates obtained in fall and/or winter to determine starting difficulty of spring assessments;
- embed GMAP field test items randomly across field test positions;
- recalibrate all MAP items to build the GMAP scale;
- enable districts to allow students to pause tests and resume on the same day or the next day;
- provide sample items months before the field test; and
- have the field test deliver linked RIT scores while collecting sufficient data for building the GMAP summative scale.

NWEA has three sets of items: (1) items that have RIT parameters, which are used to produce linked RIT scores; (2) NWEA items that come from a summative item pool and that are not on the RIT scale, and (3) newly developed items, created to measure Georgia standards that are not covered by existing items. All items have been aligned to the Georgia standards, and existing IRT parameters are being used as if they are operational for adaptive simulation purposes. All items will be calibrated based on field test data, at which point previous statistics (where available) will not be used. Existing IRT statistics are just being used to drive the adaptivity. NWEA plans to vary the positions of passages and items in the field test to analyze potential fatigue effects and item position effects. NWEA examined the stability of theta estimates for a 30-item MAP Growth test. Simulation results show good stability in total score after 30 items. NWEA will provide previews of the technology-enhanced item types and sample reports. Independent alignment will be conducted in summer 2022 or 2023.

The RIT scale is used to measure within-year growth (spring-to-spring, winter-to-spring, fall-to-spring). Instructional feedback is available via the learning continuum. GMAP is most interested in using the RIT score to see if growth targets are met. There is also the use of RIT scores (or other nationally normed assessments) to classify students into gifted programs). Maintaining the RIT scale adds value to the assessment system for score users. It also provides a continuum from K–2 through 3–8 and beyond. This will eliminate a test, so that more testing is not needed for gifted programs or other purposes.

Teachers will use the end-of-grade assessment to understand student performance in terms of the state's content standards. The norm-referenced score provides an additional interpretation about how a student is doing in relation to the nation. The two scores provide answers to different questions. It's easier for parents to think about growth on a scale that increases from grade to grade. Milestones doesn't have this feature, and Georgia has struggled to provide meaningful norm-referenced scores that parents understand how to differentiate from the criterion-referenced score. The MAP Growth items used in GMAP are aligned to the Georgia Standards of Excellence (GSEs).

The TAC noted that having sample items outside of the field test forms is acceptable. However, they should be provided in the same platform. Otherwise, the items might function differently or look different. The TAC also noted that a survey to detect student levels of effort or motivation effects might be helpful. It will be interesting to see how different the original item statistics are from the statistics that are obtained from the upcoming GMAP administration. The populations of students who took the items are different demographically and in terms of achievement levels. NWEA is cautiously optimistic, but invariance probably will not hold across the board. The MAP Growth items

have very stable statistics, and can be used to generate the RIT scores without concern. RIT items will not be recalibrated.

Both RIT-linked and GMAP scores will be produced on a single score report. The TAC asked if the information provided to teachers via the RIT scores and via GMAP provide confusing or conflicting messages. GMAP noted that there may be differences, but the RIT scores will be very similar to the RIT scores provided via the MAP Growth assessment, which teachers are familiar with. Teachers are also familiar with the GSEs, so the GMAP scores, which measure the GSEs, will also be somewhat familiar. By 2022–23, GMAP will have score reports that can be compared to see how interpretations might differ. The TAC mentioned that consequential validity will be important to look at in terms of the score interpretations of the two score reports and the decisions that are made. TAC suggested getting people's reactions to the two scores and determining whether both scores should be included for all users or just district-level users.

Range Achievement Level Descriptors

During this section of the meeting, NWEA described the work that has been conducted, to date, on the process used to adapt the GSEs to Range Achievement Level Descriptors (RALDs) for a computer-adaptive assessment. These RALDs are at the standard or substandard level for all content areas, and all represent on-grade-level content. GMAP has expanded the substandards to a finer-grained level than in the Milestones ALDs: some standards have been broken down into smaller "chunks."

GMAP will analyze data to determine whether these levels are supported empirically. These levels incorporated Georgia educator and content advisory feedback. However, if data do not support the fine-grained distinctions, the RALDs will be collapsed to a higher level. The intent is to provide more instructionally useful information throughout the year. Grades 3–8 math, ELA, and science RALDs have been completed. The current plan is to expand the process to high school.

The TAC noted that the level of detail in the GMAP RALDs may be more detail than necessary, especially given that Milestones is not at this detailed level. However, this level of detail would be helpful to item writers. NWEA is currently using this information for pool analysis and item writing; careful consideration would be needed to determine whether it could be used for reporting purposes. The TAC has an overall concern that going to a finer grain level for the RALDs may actually make demonstrating comparability to Milestones harder. The test specifications for Milestones provide the basis for alignment. The CAT algorithm will not need to select items at specific levels or substandards. To have the RALDs at this level and the blueprint at another might lead to misalignments. The TAC was also concerned that GMAP moved items to different domains because of places where NWEA felt that the Milestones RALDs had inconsistencies. This could also contribute to misalignments if it is a pervasive issue, especially given how items roll up to domain subscores. NWEA noted that by keeping the inconsistencies in the Milestones RALDs, GMAP may actually be penalized during the item-to-standards alignment process. The TAC asked for proof that finer-grained descriptions are instructionally useful. The TAC did note that once the GMAP assessment is aligned to a higher level of content, it will be challenging to evaluate the assessment at a finer grain level; if the assessment is aligned at a lower level, it is easier to roll up alignments to a higher level, if needed. It was noted that the GA standards will be updated and changes will need to be incorporated into the GMAP plan.

Alignment Study

In the last meeting, the TAC requested additional information on GMAP's first alignment study. During this presentation, NWEA provided an overview of a bank analysis that was conducted by EdMetric. This was a preliminary alignment study; an independent alignment study is planned after the first operational administration. RALDs were the focus of this exploratory alignment study. Anne Davidson from EdMetric presented the results of the study. An item-descriptor matching method was used, including ordered item booklets that were sorted by both content standards and item difficulty within subject and grade. The process included a content alignment rating, a DOK rating, and, finally, an RALD rating. The first two steps are very consistent with the traditional content alignment study, whereas the RALD rating is a novel approach. Results indicate that there are items in the bank that may measure a GSE, but there are not RALDs that match to those items. Changes to the RALDs could remedy this. Rater agreement was very high. Most items fall into DOK 1 or DOK 2, and RALD results indicated potential locations where additional items could be developed to increase the coverage of the GSEs in the GMAP item pool.

The TAC noted that the item-descriptor method is a standard setting method, not an alignment method. The TAC asked for clarification on the rating process. Anne explained that the on-grade GSEs and OIBs were provided to subject-matter experts (SMEs) to facilitate the alignment process. SMEs were also provided with adjacent below- and above-grade GSEs. Items were then compared to these GSEs. SMEs identified which content standard the item aligned best to, even if it was an off-grade-level standard. The TAC supported the ordering of items by content but was not sure that ordering by difficulty was necessary. Overall, the TAC felt that the study was interesting but not necessarily the most relevant evidence for comparability between GMAP and Milestones. The final GMAP item pool will be an amalgamated item pool that includes previous MAP items, newly written items, and other NWEA-owned summative items. Collectively, the complete GMAP item pool will align to the full range of the GSEs. This alignment study covers a portion of the GMAP item pool; future alignment studies will include a representative sample of the complete GMAP item pool.

Design of the Through-Year CAT

NWEA has performed many CAT simulations in the past year to evaluate different CAT designs. During this presentation, NWEA described its proposed CAT design, how it can be configured, and what kinds of information it can produce. NWEA sought the TAC's feedback on the following questions:

1. What types of evidence would you look for when implementing a new innovative CAT design?
2. What are the strengths and possible weaknesses of this CAT design? What recommendations might address the weaknesses?

NWEA described its goal with the CAT design as maximizing efficiency and actionable information. The design includes a modified shadow CAT approach with a weighted penalty model to create a student-specific form. Items selected for each student are based on the updated student ability estimate as the student moves through the test, along with the blueprint requirements. Early on, if the student is struggling, the engine can identify supporting off-grade skills to provide diagnostic information. There are many constraints in the system, including DOK and standards. The

constraints ensure that every student receives coverage of the standards on their assessments. NWEA described a flow chart illustrating each decision point in the CAT design.

A proof-of-concept test produced reliable scores with 27 items. In the second part of the assessment, students can be routed off grade, if necessary, to pinpoint strengths and weaknesses. Blueprints proportional to the Milestones blueprint may have some difficulties for very small domains, because the domains will include even fewer items. The engine has a lot of flexibility, but the constraints must be prioritized. The current method uses a fixed-length, rather than variable-length, CAT.

The TAC had positive feedback on the CAT model. The TAC asked how blueprint coverages ensured. NWEA explained that the first section of the adaptive assessment provides a proportional representation of the blueprint. The TAC expressed concern that there were not enough high-DOK items in the pool. Item development has focused on filling those gaps. The TAC noted that Milestones does have DOK targets, and asked whether these targets could be added to the CAT. NWEA indicated that this is definitely possible. The TAC wanted to know what NWEA is planning and which constraints they recommend moving forward with. NWEA plans to run simulations soon to understand how the constraints interact with the current item pool and will present this information to the TAC at the next meeting. The TAC encouraged NWEA to think very flexibly about all aspects of the CAT and to consider the proportion of students who received an assessment that met the Milestones blueprint in terms of content and cognitive complexity. The TAC mentioned that having enough items to provide the data required for reporting is important. The TAC requested to have sample score reports to understand how many items will be needed. The TAC also recommended exploring, through simulations and focus groups, how much flexibility in terms of test length and other features is acceptable if there are real benefits in terms of score precision. Having the ability to include so many different constraints and guidelines is great, but results still need to be interpretable by users.

The TAC mentioned that it is important to verify that the score precision for subscores/diagnostic categories is sufficiently high for reporting purposes, and to ensure that the CAT can satisfy the requirements of the federal IADA and, at the same time, supports the theory of action. Items should measure a full range of the content, rather than there just being enough items within a domain to provide a subscore. The consortium can use the distribution of ability in the Georgia student population to see how constraints in the CAT model play out. There are only so many constraints that can be supported, but GMAP should attempt to push the boundaries. The TAC really wants to see how the students are funneled through the item pool and what the content representation and score precision look like for a representative sample of student assessments. The TAC recommended looking at the balance of items between the on-grade and diagnostic sections: How does that differ by grade, ability level, subject, etc.? Also, what percent of students receive below-grade items? Above-grade items? Although it is not the most critical piece of evidence, looking at the item response time will be critical. The test could be timed, or not, depending on client requirements.

The TAC mentioned that the blueprint coverage could only be based on the items that contribute to the summative score. If GMAP moves forward with including only the results from the final assessment in the summative score, the content/blueprint coverage should focus on the final

assessment. The TAC supported NWEA's proposal to use previous assessments to inform the starting difficulty of subsequent tests.

Timeline and Next Steps

In the last meeting, the TAC requested additional information on GMAP's theory of action, score reporting, and professional learning plans. A presentation on these topics was planned for this meeting but was postponed due to time constraints.

The primary objective during the next TAC meeting (December 2021) will be to show the TAC the progress that has been made on comparability. Comparability evidence artifacts or descriptions, aligned to the requirements of the comparability guidelines, should be provided as pre-meeting materials to the TAC. The TAC will not provide a thorough review of a substantial amount of documentation prior to the December meeting, but providing as much documentation to the TAC as possible, along with an indication of whether the documentation is in draft format or finalized, will help the TAC understand the consortium's progress and technical assistance needs for 2022.

For areas of the checklist where evidence/artifacts have not yet been created, the timeline and process for assembling those pieces should be described. It will be good to show the TAC how far the consortium has been able to come in the past two years, despite the pandemic; how delays have impacted timelines; and a high-level schedule of the upcoming three years. For example, when does it look possible to implement in lieu of Milestones for grades 3–8 ELA and math? What about science and social Studies? What about high school? Implementing the full set of assessments in the same year is not necessary, but there should be a long-term plan and timeline to fully replace Milestones.

The TAC is also interested in the consortium's theory of learning and theory of action. If there are areas of the checklist where the consortium differs from Milestones, is there evidence that those differences are improvements?

Following is a list of topics in which the TAC has expressed interest:

- Theory of learning/theory of action
- Summative score determination (including score banking decision)
- Score reporting
- CAT simulation results
- Accessibility and accommodations
- Professional learning plans

These and other TAC topics should be prioritized based on how relevant they are to the comparability guidelines and how soon answers are needed, based on the consortium's timelines.

Georgia Innovative Assessment Pilot Program

JULY 2021 TECHNICAL ASSISTANCE COMMITTEE MEETING

Putnam County Consortium

Mariann Lemke
Sonya Powers
Assessment Research & Innovation @WestEd | csaa.wested.org

GEORGIA INNOVATIVE ASSESSMENT PILOT PROGRAM

JULY 2021 TECHNICAL ASSISTANCE COMMITTEE REPORT FOR PUTNAM COUNTY CONSORTIUM

INTRODUCTION

The Georgia Innovative Assessment Pilot Program (IAPP) Technical Advisory Committee (TAC) met on July 8, 2021, 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. The agenda included two main topics:

- a review of comparability requirements and associated discussion of their specific application to the Navvy assessments; and
- an update on Navvy's implementation.

This report provides an overview of each topic and a description of the resulting key takeaways and action items from the meeting.

COMPARABILITY REQUIREMENTS CHECKLIST

To begin the meeting, WestEd staff provided an overview of the comparability evidence that the consortium will be required to provide to the state. Examples of relevant evidence are described in a template that will be provided to Putnam. Evidence is required in several main categories, as described in the following sections.

Alignment and Comparability

Consortium assessments must demonstrate that:

- assessments and items are aligned to the Georgia standards,
- assessments match the depth and breadth of the Georgia standards,
- students can be classified into at least four achievement levels representing the same knowledge and skills that current Milestones assessment achievement level descriptors (ALDs) provide,
- summative classifications of students are consistent across Milestones and innovative assessments (for all students, subgroups of students, content areas, and assessments),
- those who participate in the innovative assessment are representative of the state in terms of demographic composition and achievement, and
- there is a plan for conducting annual comparability analyses between the innovative assessment and Georgia Milestones throughout the remainder of the IADA period.

To meet these criteria, the consortium should present an independent alignment study including information similar to that provided in previous Milestones reports. Four types of alignment should be included: balance of complexity, depth and range of knowledge, and categorical concurrence. Note that conducting an alignment study of all items is not necessary (though every grade level should be included). A sampling approach that provides strong evidence that the items and tests that students actually encountered on a consortium assessment are aligned (for example, by selecting a sample of students across proficiency levels and checking alignment for those students' tests) can suffice. Note also that the state is updating its standards. New math standards will become operational in 2023–24 and ELA in 2024–25, so new evidence of alignment will be needed after the new standards become operational.

The consortium must also demonstrate that it has achievement levels that correspond to the current Milestones ALDs. Direct adoption of Georgia's ALDs can satisfy this criterion, though other ALDs may be used with evidence of their alignment to the existing ALDs. The consortium must show evidence that students at each of the Milestones ALD levels have the skills and knowledge described in those ALDs. For example, if the Milestones ALD describes proficiency as being able to use place-value relationships to round numbers, the consortium should demonstrate that students placed into that performance level on the innovative assessment also demonstrate those skills.

The consortium must also provide a report on how classification into its achievement levels compares to classifications on the Milestones assessment. Only on-grade-level items should be used to classify students into performance levels. It is possible that new tests may provide different results for good reasons, based on the design of the assessment or the approach to scoring; the consortium should be prepared to fully explain and justify why differences may occur. The consortium should be sure to describe not just how many students are at each level but the degree to which students are consistently classified by the two assessments. Because end-of-course assessments contribute 20% to course grades, the consortium should also provide evidence of its approach to using its scores for grades and the comparability of those grades to the grade conversion score (GCS) method used with the Milestones assessments.

Consortium documentation should also include descriptive analyses of its participating populations of students, compared to the state, with description of weighting methods or other mechanisms for generalizing sample results to the state, as relevant. All state-reported subgroups of students should be included, as well as a description of groups based on achievement.

Beyond initial comparability analyses based on students taking both the consortium assessments and the Milestones tests, the consortium must provide a plan to conduct annual comparability analyses for the remainder of the IADA period. This plan need not include testing of all students, but, rather, should include a sample of grade bands (or grade bands/students), so that each grade band includes an innovative assessment and the state assessment (see IADA final regulations, pp. 28–29).

Technical Quality

The consortium must also provide evidence of the technical quality of its assessments, demonstrating:

- work with experts to ensure quality,
- reliability and validity of the assessments,

- how the assessment provides information across the full performance continuum for students,
- availability of individual and aggregate reports and the timeliness and interpretability of these reports for stakeholders,
- how principles of universal design for learning were incorporated into the assessment design, and
- a plan to maintain the item bank and the integrity of the score scale over time.

To meet these criteria, the consortium should provide background information (e.g., names, CVs) of TAC members and agendas of meetings aimed at discussing technical quality of the assessments.

The consortium should also present evidence of validity that matches the categories in the *Standards for Educational and Psychological Testing.* Not all evidence (e.g., consequential validity) may be available immediately, but the consortium should describe its plan to gather this information over time. Consideration of what validity evidence can be provided without testing, what can be gathered during piloting, and what must be gathered once an innovative assessment is fully operational may be useful.

The consortium must provide reliability evidence for the summative scores, subscores, and achievement levels generated from the innovative assessment, consistent with national standards and the Georgia Milestones. For example, evidence might include test-subtest reliability (again, including only on-grade-level items). Decision consistency and accuracy values should be similar to those reported for Georgia Milestones.

Data showing the distribution of scores, to demonstrate how the assessment provides information across the performance continuum, should also be presented. These data could include analyses of test information functions or other analytics, or other types of information such as cognitive lab data and test blueprints indicating depth-of-knowledge ranges.

The consortium should provide examples of its student and aggregate-level reports (such as classroom, school, consortium, and even state-level reports). These reports should be accompanied by evidence that stakeholders can use these reports to make valid interpretations about student performance, such as data drawn from focus groups of a variety of stakeholders representing report consumers, data from A/B tests, or other data.

Innovative assessment reporting timelines must describe when and how stakeholders receive results of the assessment, demonstrating that these results are provided in a timely manner. Final results for accountability must be provided at least in the same timeframe in which the current Georgia Milestones assessment final results are available.

The consortium should also provide a description of how its assessments incorporated principles of universal design for learning in test development, as well as how scales and item banks will be maintained over time (e.g., how parameter drift will be managed).

Accessibility and Accommodations

All students who currently participate in Georgia Milestones must be able to participate in the innovative assessment in order to use the innovative assessment in lieu of Georgia Milestones, including students with disabilities and English learners (except students with the most severe cognitive disabilities, who may participate in an alternate assessment).

A crosswalk of accessibility and accommodation features available on Georgia Milestones and available on the innovative assessment should be provided such that it is possible to see, at a glance, whether all of the accessibility and accommodation features will be available, and, if not, how students will be validly assessed using an alternative accessibility mechanism. Any differences in the ways that accessibility or accommodation features work in the innovative assessment, compared to Georgia Milestones, should be indicated.

Accessibility features and accommodations must allow students to participate in alignment with their IEPs or English learning plans and comply with relevant federal laws such as the Individuals with Disabilities Education Act (IDEA). The consortium should provide a participation report that shows that all students are participating as required.

The consortium need not have all accommodations available in order for the innovative assessment to be approved for use in lieu of the Georgia Milestones, but must have a specific and feasible plan to provide all needed accommodations when assessments are administered. For example, the consortium need not have Braille forms ready at the time that evidence of comparability is being reviewed, but must have a well-described plan to produce Braille forms prior to administration, that demonstrates the vendor's capacity to produce them (historical evidence of how they have been produced in the manner described).

Test Administration and Security

The consortium must demonstrate that it has plans in place to ensure standardized administrations, such as training and manuals, and processes to prevent and/or document testing irregularities and protect test security and student data. In addition, the Georgia Office of State Assessment will monitor consortium test administrations, and monitoring reports should be included in evidence for this criterion. Other evidence would be sample irregularity reports, results of analytical analyses aimed at discovering cheating, auditing procedures, and procedures to handle irregularities or test security violations.

The consortium should keep in mind that standardization processes are intended to promote the validity and comparability of the scores, but the consortium need not compromise features of the assessments that make them innovative. As an example, using many different types of accommodations reduces the standardization of administration, but is necessary to ensure validity of the scores.

Stakeholder Engagement

The consortium should provide evidence that assessments were developed 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. Evidence might include letters of support or agendas from meetings where assessments were discussed, along with participant lists.

The consortium should also document how it has worked with schools and districts to interpret results and communicate with stakeholders such as parents, students, and community members (i.e., how the consortium has worked to develop assessment literacy). Evidence might include training agendas and presentations, meeting agendas, assessment guides, score interpretation guides, data on stakeholder participation in training for test administration or score interpretation, or stakeholder survey or focus group data.

Accountability

Georgia's accountability requirements must be met with use of any innovative assessment. In addition to the need to provide a summative score, these requirements also include providing measures for the College and Career Ready Performance Index (CCRPI).

The consortium should demonstrate that it uniquely identifies students within and across years so that students' assessment data, schools, districts, demographic information, etc., can be used for accountability purposes. Data layouts and timelines should be provided. Evidence must also be provided that the percentage of students assessed is at least as high as the percentages observed on Milestones prior to the start of the innovative pilots, overall, as well as for all federally required student demographic subgroups.

The consortium must describe how it will produce a single summative score. If there is more than one administration during the academic year (e.g., a through-year model), the consortium should specify which administrations contribute to the summative score and how scores are combined. This description should provide a clear rationale for the calculation of the summative score.

As noted, the consortium must also show how its assessment data can be used for a variety of CCRPI purposes, including providing measures for the Content Mastery and Closing Gaps components of the index, growth measures for the Progress component, and literacy measures for the Readiness component. These measures do not need to be strictly comparable to, or use the same methods as, the Georgia Milestones, but evidence must be provided that justifies the proposed approach.

Conflict of Interest

The consortium must provide assurances that there are no conflicts of interest (financial or otherwise) for parties participating in the pilot program, and that all local procurement rules are being followed. No new evidence is needed unless there have been changes since initial assurances were made at the award of the innovative assessment grants.

TAC Discussion and Recommendations

The TAC noted several aspects of the comparability requirements that the consortium will need to carefully consider, including the following:

Participation

Given the ongoing nature of the innovative assessments, how is participation defined? TAC members also raised the issue of student mobility and requested that the consortium consider how to handle situations where students transfer in late in the school year and may not have participated in earlier assessments. How can a summative score be produced in these situations? The consortium may need to consider business rules such as the "attemptedness" rules that the Milestones uses to determine what counts as participation, and what is needed to be able to make a judgment about student proficiency. One way to think about this might be to focus on "culminating" standards that incorporate prior standards and skills from within the grade.

Retention of Learning

TAC members also noted that the current Milestones exams assume that students will retain information they may have learned earlier in the year and be able to demonstrate it on an end-of-year test. Innovative assessments may use a different model of learning, where scores represent an accumulation of information about learning from different points, rather than from one moment in time. Description of what the final scores reflect, and how that may be the same as or different from the Milestones model, will be important.

Multiple Opportunities

Because the consortium's approach allows students to attempt to demonstrate mastery of standards up to three times, the vendor should be sure to analyze the use of multiple attempts and thoroughly document how and when multiple attempts are incorporated into reporting—how they are used, when, on which reports, and how their use impacts results. The vendor noted that its item selection algorithm prioritizes depth and breadth of standards first, then new items, so it is also possible that students could see the same items over time. These situations should also be documented.

Use of Assessment for Accountability

TAC members noted that the system is trying to serve multiple purposes: to provide useful information for feedback and instruction, and, ultimately, to provide measures that can be used for accountability. While the focus now may be on feedback and instruction, behavior and use of the data may change once the assessment is being used in lieu of the Milestones for accountability purposes. The consortium should consider how to gather information on the use of data, both before and after administration of Navvy in lieu of Milestones, to report on consequential validity.

Ongoing Nature of Reporting

Because the assessment system aims to provide real-time information to inform instruction, users have data about student performance at all times. TAC members noted that there is potential for misuse of the data if users don't understand what is included and what it represents, and try to make summary judgments before assessment is really complete. The TAC suggested that the consortium consider how and when to report "final" data, particularly at aggregate levels such as the district level or even the state level, so that appropriate interpretations of the data can be made. Such an approach may be especially important if summative classifications are potentially available on an ongoing basis.

Pacing and Coverage

Different classrooms may provide instruction at different speeds, even if all are following a common pacing guide. With any type of high-stakes assessment, teachers may rush to cover as much of the expected content of the assessment as possible prior to administration. This situation may be exacerbated when assessments don't just take place at the end of the year, but are spread out throughout the school year. The consortium should consider how to balance the need to allow for variability in assessment administration windows with the need to maintain some standardization. It is also important to help consortium members avoid situations where schools or teachers are rushing not just to cover content but also to administer multiple assessments toward the end of the year. Training and handbooks may be an important element to address these types of concerns.

Integration of Standards

TAC members asked about integration of standards. Navvy's current design assesses individual standards in isolation, though it was pointed out that some standards include knowledge and skills from prior standards (and that standards are not necessarily taught in isolation, even if they are assessed in that manner). Though this is not necessarily included in the comparability criteria, the TAC suggested being sure to describe this aspect of Navvy's learning and assessment model when discussing interpretation of results.

Update on Consortium Assessment System

Goals and Features of Assessments

A key goal of the Navvy assessment system is to provide validity and reliability around standards-based reporting. The Navvy assessments are intended to inform teaching and to guide learning by accurately identifying what learning has taken place and what learning needs more support. An aim of the current work is to leverage Navvy's assessment data for everyday use in monitoring student learning as well as for accountability purposes.

Hallmark features of the Navvy system are the real-time reports that provide an at-a-glance update on student mastery of standards. The design is intended to be diagnostic at the standards level. Teachers determine when to give assessments, based on their instructional pacing, and information on mastery is updated as soon as it is available. Students may take assessments up to three times;

this design is aimed at helping create a growth mindset in which students are not simply "not proficient," but, rather, are "not yet proficient," and will have additional opportunities to demonstrate their learning. Teachers cannot see the items that contribute to the accountability assessments, but they can see the items for the practice assessments. The TAC asked whether students have the same awareness of Navvy as an assessment, compared to Milestones. Students do know that it is an assessment event, not just part of a learning management system. Teachers do not typically use Navvy for grades, especially in elementary school, though this may shift at middle school and high school.

Sample Reports

The consortium also showed sample student and teacher dashboard reports, which provide a quick way for the user to see each standard and whether the student has demonstrated mastery of that standard. Reports can be extended to look at performance over years or across classrooms as well.

Summative Score Calculations

The consortium offered several initial ideas on summative scoring; it is evaluating multiple approaches using the data collected in 2019-20 and 2020-21. An initial idea is to calculate the percentage of standards mastered as the summative score. Thresholds could be placed on the percentage metric to delineate the achievement levels. By default, everyone would start in the lowest category and move up toward the highest as they test and pass more standards. They could then see where they are throughout the year in terms of achievement level/accountability metric. Another approach could use a weighted percentage of standards mastered, using the Milestones blueprint, to have the number of standards by domain for Milestones drive the Navvy weights.

Initial Data on Reliability and Comparisons to MAP

Navvy showed some preliminary data from 2020–21, including the base rates of competency mastery in fourth grade math, using only the first attempt. Reliability at the standard level is almost always 0.8 or above (all above 0.7). Each standard is measured by 6–9 items. Item discrimination analyses also seemed to be within industry standard ranges.

The consortium also provided some more-detailed results from an analysis of MAP and Navvy scores in math. The analysis showed that there are several standards profiles from Navvy that correspond to the same MAP Growth scores—that is, students' scores may be exactly the same on MAP subscales, but the pattern of their standards mastery as demonstrated in Navvy can be quite different. Scores between the Navvy and MAP scales are correlated at about 0.5. The TAC noted that the MAP-to-Navvy comparison should be replicated with scores from Milestones, which could provide comparability evidence. The more of the state's variability that is included in the analysis, the more informative it will be. The TAC suggested identifying real outliers and trying to explain why the differences are happening.

Potential Timelines and Next Steps

Putnam described some timeline options, along with some gaps between where the program is now and what will be needed to satisfy the comparability checklist. One option is to try to get ready to be operational by 2022–23, with the TAC approving use in lieu of Milestones in summer 2022. Comparability evidence would be provided to the TAC beginning with the December 2021 TAC meeting, using 2020–21 data. Use of the 2020-21 data may be challenging given participation and administration constraints due to COVID-19. TAC members noted that confidence in the Milestones scores and confidence in Navvy scores have to be high in order to make the comparability argument. Alternative approaches (e.g., Andrew Ho's metrics) might enable comparisons of the 2020–21 data to previous, more trustworthy years.

The goal would be to then add 2021–22 data and submit data in an agreed-upon format in summer 2022 so that the consortium could begin assessing in lieu of Milestones in Fall 2022.

One outstanding question is if there might be additional federal flexibility, such as extensions to states' IADA periods or waivers, to support this project. A two-year extension from the federal government might be acceptable; however, the Putnam Consortium districts are eager to move the timeline up.

TAC review of comparability materials should be staggered, as reviewing all of the documentation during a single one-day meeting won't be possible. Information could also be staggered to GaDOE. A next step is to review the timeline more thoroughly and propose a method to deliver materials in advance of the December meeting so that the meeting time can be used efficiently to gather TAC feedback.

The primary objective during the next TAC meeting (December 2021) will be to show the TAC the progress that has been made on comparability. Comparability evidence artifacts or descriptions, aligned to the requirements of the comparability guidelines, should be provided as pre-meeting materials to the TAC. The TAC will not provide a thorough review of a substantial amount of documentation prior to the December meeting, but providing as much documentation to the TAC as possible, along with an indication of whether the documentation is in draft format or finalized, will help the TAC understand the consortium's progress and technical assistance needs for 2022.

For areas of the checklist where evidence/artifacts have not yet been created, the timeline and process for assembling those pieces should be described. It will be good to show the TAC how far the consortium has been able to come in the past two years, despite the pandemic; how delays have impacted timelines; and a high-level schedule of the upcoming three years. For example, when does it look possible to implement in lieu of Milestones for grades 3–8 ELA and math? What about science and social studies? What about high school? Implementing the full set of assessments in the same year is not necessary, but there should be a long-term plan and timeline to fully replace Milestones.

The TAC is also interested in the consortium's theory of learning and theory of action. If there are areas of the checklist where the consortium differs from Milestones, is there evidence that those differences are improvements?

Following is a list of topics in which the TAC has expressed interest:

- Theory of learning/theory of action
- Additional results from 2019–20 or 2020–21
- Summative score determination
- Assessment plan for students who are not in the district for the full year
- Plan for the literacy CCRPI measure
- Accessibility and accommodations

These and other TAC topics should be prioritized based on how relevant they are to the comparability guidelines and how soon answers are needed, based on the consortium's timelines.

Appendix 2

Georgia Innovative Assessment Pilot Program 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)

Appendix 3

Georgia Innovative Assessment
Pilot Program Comparability Guidelines

Please specify the end-of-grade and/or end-of-course assessments for which evidence is being provided for the innovative assessment.

ELA	MATHEMATICS	SCIENCE	SOCIAL STUDIES
☐ **Grade 3**	☐ Grade 3		
☐ **Grade 4**	☐ Grade 4		
☐ **Grade 5**	☐ Grade 5	☐ Grade 5	
☐ **Grade 6**	☐ Grade 6		
☐ **Grade 7**	☐ Grade 7		
☐ **Grade 8**	☐ Grade 8	☐ Grade 8 ☐ HS Physical Science (Grade 8)	☐ Grade 8
☐ **American Literature and Composition**	☐ Algebra I/Coordinate Algebra	☐ Biology	☐ U.S. History

For each of the assessments selected in the table above, evidence will need to be submitted for each of the criteria in the seven categories below (alignment and comparability, technical quality, accessibility and accommodations, test administration and security, stakeholder engagement, accountability, and conflict of interest). Note that all evidence submitted should be based on grade-level items only. Off-grade items can be included on assessments but cannot be included in the evidence required below.

1 Alignment & Comparability

	Criteria	Yes	No	Examples of Relevant Evidence	Evidence Documents* (pages)	Commentary (Optional)
1	Do you have an independent alignment study between the innovative assessment and the Georgia academic content standards (GSEs) for all grades, content areas, and courses? Note: The revised mathematics GSEs are expected to be operational for the 2023-2024 school year and the revised ELA GSEs are expected to be operational for the 2024-2025 school year.	☐	☐	Alignment study report	<*Consortium A Alignment Report 2022.docx*> (*1-35*)	
2	Does the alignment study indicate that the innovative assessment adequately reflects Georgia academic content standards for all grades, content areas, and courses in terms of categorical concurrence, balance of representation, depth of knowledge, and range of knowledge? Note: If the innovative assessment is computer adaptive, documentation should demonstrate procedures that ensure the item pool and content constraints result in good alignment at the student level across all ability levels.	☐	☐	Alignment study report • Similar to alignment of Georgia Milestones Test blueprints indicating depth of knowledge ranges/cognitive complexity levels Item and passage specifications Item selection procedures	<*Consortium A Alignment Report 2022.docx*> (*32-33*)	
3	Does the innovative assessment classify students into four achievement levels that are consistent (representing similar levels of knowledge and skill) with those reported for Georgia Milestones? Note: Direct adoption of Georgia's ALDs is recommended to satisfy this criterion. If	☐	☐	Achievement level descriptors	<*Consortium A Statewide Performance SY21-22.pdf*> (*2*)	

	other ALDs are used, they must be justified and the alignment to the Georgia ALDs evaluated.					
4	Are summative classifications of students into the four achievement levels consistent between the innovative assessment and Georgia Milestones for all students and for all subgroups of students across all grades, content areas, and courses? Note: A standard setting is not expected, rather, empirical methods can be used to set cut scores on the innovative assessment that results in consistent student classifications into achievement levels. If the innovative assessment contains any off-grade level items, achievement level classification should be determined using only items that measure on-grade level standards (i.e., the grade in which the student is enrolled) and uses that determination for reporting and accountability. Consortia should also be aware that end-of-course assessments contribute 20% to course grades. The grade conversion score (GCS) is tied to the scale score cuts for Developing Learner and Proficient Learner. Specifically, for Georgia Milestones, the GCS ranges from 0 to 100. GCS=0 is set to the LOSS, GCS=100 is set to the HOSS. GCS=68, 80, and 92 are set to the scale cuts between achievement levels (1/2; 2/3; 3/4). A linear transformation is applied to obtain the GCS values between the points above.	☐	☐	Classification consistency methods report, including achievement level classification consistency values and 4 x 4 contingency table for all grades, content areas, and courses for all students and all subgroups of students: • Exact Agreement (>0.7) • Exact + Adjacent Agreement (>0.9) • Quadratic Weighted Kappa (>0.85) The report or associated evidence should document, as applicable: methodology, calibration model(s), assumption check results, reliability, mean/range item difficulty, distribution of item types across the scale, student sample exclusions and impact of exclusions, consistency of results by demographic subgroups, comparability of administration conditions (e.g., speededness, format). The classification consistency report should also include an analysis of how comparable student grades are likely to be for end-of-course assessments given the GCS method.	<*Consortia A vs. Milestones Performance Level Classification Consistency (SY21-22).docx*> (1-30; results pages 28-31)	
5	Are the students who participate in the innovative assessment representative of	☐	☐	Table of sample vs. state demographics and achievement		

	the state in terms of demographic composition and achievement? Note: If the answer to this question is no, then provide evidence demonstrating how the sample has been weighted or adjusted to represent the state when necessary.			(include all subgroups reported in Georgia for accountability) Description of weighting methods or other mechanisms for generalizing sample results to the state.		
6	Do you have a plan for conducting annual comparability analyses between the innovative assessment and Georgia Milestones throughout the remainder of the IADA period? Note: Comparability analyses will require double testing of Georgia Milestones and the innovative assessment for a sample of grades and subjects.	☐	☐	Comparability analysis plan		

*The Evidence Documents column can either contain the file name(s) of the relevant artifact(s), or a hyperlink to the document.

2 Technical Quality

	Criteria	Yes	No	Examples of Relevant Evidence	Evidence Documents (pages)	Commentary (Optional)
1	Have you worked with experts to ensure technical quality, validity, reliability, and psychometric soundness of the innovative assessment?	☐	☐	CVs/qualifications of technical team Meeting agendas or meeting summaries (e.g., internal meetings, WestEd technical assistance meetings, TAC meeting transcripts, other consultant meetings)		
2	Have you established reliability evidence for the summative scores, subscores, and achievement levels generated from the innovative assessment consistent with nationally-recognized testing standards? Notes: For preliminary or on-demand results/scores, demonstrate the technical	☐	☐	Reliability section of the technical report (include overall reliability, subscore reliability, conditional standard errors of measurement, decision consistency, and decision accuracy)		

	evaluation procedures used to evaluate consistent reliability, including evaluation of model assumptions/parameters/scale stability. As a point of comparison, the majority of Georgia Milestones EOG and EOC assessments have reliability values of 0.9 and above. Include subscore reliability, but strict reliability criteria will not be required. Decision consistency and accuracy values should be similar to those reported for Georgia Milestones.					
3	Have you established validity evidence for the innovative assessment consistent with nationally-recognized testing standards? Note: Much of the Comparability assurances criteria also provide validity evidence. Content evidence is most critical, relations to other variables will be available through comparison to Georgia Milestones, and validity evidence should be organized around the five sources of validity evidence described in *The Standards.* Evidence of test consequences, especially as it relates to the theory of action should be provided as soon as possible.	☐	☐	Validity section of the technical report Blueprints, test specifications, alignment studies		
4	Is the innovative assessment designed to assess student achievement based on grade-level state academic content standards in terms of content and cognitive processes, including higher-order thinking skills, and to adequately measure summative student performance across the full performance continuum for all students, except students with the most significant cognitive disabilities?	☐	☐	Score distributions Test blueprints, assessment guides, or other documents indicating depth of knowledge ranges Summary of item types Item and passage specifications Cognitive labs or other studies addressing student cognitive processes Analyses of test information functions demonstrating precision across the performance continuum or other demonstration of information function		

				across the performance continuum CSEM across the scale/at the cut points Analyses (e.g., differential item functioning (DIF), differential test functioning (DTF) analyses) that identify possible bias or inconsistent interpretations of results across student groups Alignment studies		
5	Do you produce individual student score reports?	☐	☐	Example student report Score interpretation guide		
6	Do you produce aggregate score reports?	☐	☐	Example classroom, school, district, consortium reports Score interpretation guide		
7	Have you collected evidence that students, parents, educators, and school leaders are able to use your score reports to make valid score interpretations? Note: Include information about the representativeness of the sample for each stakeholder group.	☐	☐	Reports from cognitive labs, focus groups, etc.		
8	Are score reports provided in a timely manner?	☐	☐	Reporting timeline (e.g., number of days between the administration and when score users are provided with preliminary and/or final results along with activities occurring between the two milestones)		
9	Have you incorporated principles of Universal Design for Learning into your innovative assessment?	☐	☐	Test development chapter of technical report Accessibility/UDL reports		
10	Have you developed a maintenance and evaluation plan to address longitudinal scale stability, identification and mitigation of parameter drift, and bank maintenance?	☐	☐	Psychometrics, research, and evaluation section of the technical report Details on item pool		

3 Accessibility & Accommodations

All students who currently participate in Georgia Milestones must be able to participate in the innovative assessment in order to use the innovative assessment in lieu of Georgia Milestones. A crosswalk of accessibility and accommodation features available on Georgia Milestones and available on the innovative assessment should be provided such that it is possible to see at a glance whether all of the accessibility and accommodation features will be available, and if not, how students will be validly assessed using an alternative accessibility mechanism. Any differences in the way accessibility or accommodation features work in the innovative assessment as compared to Georgia Milestones should be indicated. Over time, the accessibility and accommodation features available for use on the innovative assessment should improve to reach industry best-practice.

	Criteria	Yes	No	Examples of Relevant Evidence	Evidence Documents (pages)	Commentary (Optional)
1	In participating schools, are all students, except those with the most significant cognitive disabilities, participating in the innovative assessment?	☐	☐	Participation rate report Table of sample vs. state demographics and achievement		
2	Are students with disabilities provided with appropriate accommodations as defined by their IEP/IAP?	☐	☐	Relevant sections of the accommodations manual List of available accommodations Braille and VSL materials/resources Results of analyses and/or expert review indicating that accommodations do not alter the construct (e.g., classification consistency studies, DIF studies, person fit studies)		
3	Are English learners provided with appropriate accommodations as defined by their EL/TPC?	☐	☐	Relevant sections of the accommodations manual List of available accommodations Results of analyses and/or expert review indicating that accommodations do not alter the construct (e.g., classification consistency studies, DIF studies, person fit studies)		
4	Do all provided accessibility tools and accommodations comply with all federal laws, including, but not limited to, IDEA, ADA, Section 504 of the Rehabilitation Act of 1973, Title I, ESEA, and FERPA?			Relevant sections of the accommodations manual		

4 Test Administration & Security

If some of the test administrations do not contribute to a summative score, then the test administration and security requirements could be reduced. However, items from high-stakes administrations should not also be used during low-stakes administrations.

	Criteria	Yes	No	Examples of Relevant Evidence	Evidence Documents (pages)	Commentary (Optional)
1	Has GOSA monitored your test administrations? Note: The consortia should work with GOSA and GaDOE to develop and implement a test monitoring plan.	☐	☐	Communications with GOSA GOSA audit reports		
2	Do you have policies and procedures to ensure standardized test administration?	☐	☐	Test coordinator manuals, test administration manuals, accommodations manuals, test preparation materials for students and parents, other documents provided to schools and teachers that address standardized test administration and any accessibility tools and features available for the assessments Irregularity reports Proctor/test site training certificates		
3	Are all school staff that are involved in the test administration trained on standardized procedures and test security protocols?	☐	☐	Training presentation slides, documents, agendas Student assessment handbook Administration protocols Accessibility and accommodations manual Other comprehensive test administration policy documents Proctor/test site training certificates		
4	Do you have a process for monitoring the innovative assessment administration?	☐	☐	Relevant sections of the test coordinator manual Consortium monitoring analysis/report		

5	Do you have policies and procedures to prevent testing irregularities and ensure the integrity of test results?			Relevant sections of the student assessment handbook or assessment administration protocol manual Irregularity reports Monitoring results Data forensic methods and results		
6	Do you have test security policies and procedures to protect the integrity and confidentiality of test materials, test-related data, and personally identifiable information as established by the Family Education Rights and Privacy Act (FERPA) and the Georgia Student Data Privacy, Accessibility and Transparency Act of 2016?			Relevant sections of the student assessment handbook, test administration manual		

5 Stakeholder Engagement

	Criteria	Yes	No	Examples of Relevant Evidence	Evidence Documents (pages)	Commentary (Optional)
1	Did you develop the innovative 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? Note: Consultation with these groups is required at the beginning on the project; ongoing consultation is not required.	☐	☐	Meeting schedules, meeting agendas, letters of support, meeting participants and associated demographics or background information		
2	Did you develop capacity for educators and schools and districts leaders to implement the innovative assessment, interpret results, and communicate with stakeholders?	☐	☐	Training agendas and presentations, meeting schedules, meeting agendas, other training materials, assessment guides, study/resource guides, item and scoring samplers, professional learning offerings, score interpretation		

				guide, data on stakeholder participation in training for test administration, official logs for materials distribution, stakeholder survey results		

6 Accountability

CCRPI growth, gaps, and literacy measures do not need to be strictly comparable, nor are the innovative assessments required to use the same methods that are currently used for Georgia Milestones. The methods do need to be justified and defensible.

	Criteria	Yes	No	Examples of Relevant Evidence	Evidence Documents (pages)	Commentary (Optional)
1	Do you have a process for identifying students uniquely within and across years so that students' assessment data, schools, districts, demographic information, etc. can be used for accountability purposes? Note: The consortia should work with GaDOE to develop a data layout and reporting timeline.	☐	☐	Database with unique student identifiers (e.g., Georgia Testing Identifier [GTID])		
2	Is the percentage of students (overall and by subgroup) that you assessed in the current academic year at least as high as the percentage assessed using Georgia Milestones in the year previous to the start of the pilot (i.e., 2018-2019)?	☐	☐	Participation rate report		
3	Do you produce a single, summative score for every student? Note: If there is more than one administration during the academic year (e.g., a through-year model), specify which administrations contribute to the summative score and how scores are combined. This description should provide a	☐	☐	Scoring section of the technical report		

	clear rationale for the calculation of the summative score.					
4	Do you produce a growth measure that can be used for the CCRPI Progress component?	☐	☐	Growth measures section of the technical report		
5	Do you produce an achievement measure that can be used for the CCRPI Content Mastery and Closing Gaps components (alignment to Beginning, Developing, Proficient, and Distinguished Learner achievement levels)?			Scoring section of the technical report		
6	Do you produce a literacy (Lexile) measure that can be used for the CCRPI Readiness component? Note: Classification consistency should be demonstrated for **two** designations: Reading Status as reported for Georgia Milestones and the literacy indicator as reported for CCRPI.	☐	☐	Classification consistency methods report		
7	Do you produce 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, economically disadvantaged)?	☐	☐	Consortium summary report		

7 Conflict of Interest

	Criteria	Yes	No	Examples of Relevant Evidence	Evidence Documents (pages)	Commentary (Optional)
1	Is there a conflict of interest (financial or otherwise) for the interested parties participating in the pilot program?	☐	☐	N/A	N/A	
2	Do all activities that are related to this pilot abide by local procurement requirements?	☐	☐	N/A	N/A	