

Putnam Consortium’s Use of the Innovative Navy Assessment System’s Literacy Measure

Georgia’s school accountability system is the College and Career Ready Performance Index (CCRPI). One component of the CCRPI is Readiness. The Readiness component of CCRPI includes three indicators for elementary and middle schools (literacy, student attendance, beyond the core) and five indicators for high schools (literacy, student attendance, accelerated enrollment, pathway completion, college and career readiness). Only the literacy indicator utilizes data from the statewide assessment system, Georgia Milestones. The remaining indicators are not impacted by the implementation of the Navy innovative assessment system being implemented by the Putnam Consortium.

For the literacy indicator, the Georgia Milestones ELA assessment produces a Lexile score. The literacy indicator calculates the percent of students demonstrating reading comprehension at or above the midpoint of the College & Career Ready “Stretch” Lexile Band for each grade level. Thus, the literacy indicator utilized by the Georgia Milestones enters the CCRPI calculation as a binary indicator with two levels: (1) reading comprehension at or above grade level midpoint and (2) reading comprehension below grade level midpoint.

The Putnam Consortium will analogously utilize a binary literacy measure produced by the Navy assessment system. The Navy assessment system’s literacy measure will also produce a dichotomous appraisal of reading comprehension with two levels: (1) reading comprehension at or above grade level and (2) reading comprehension below grade level. More details on this measure are provided in the “Navy Assessment System Literacy Measure” section below.

The Putnam Consortium’s literacy measure produced by the Navy assessment system will be used in lieu of the measure produced by the Georgia Milestones assessment in the literacy indicator.

The innovative assessment technical advisory committee (or Georgia’s Accountability Technical Advisory Committee, as determined by the GaDOE) will review the literacy measure produced by the Navy innovative assessment system and its relationship to the literacy measure produced by Georgia Milestones to determine the extent to which the literacy measures are comparable. The TAC will also be asked to provide recommendations for next steps based on those comparisons. More details on comparability analyses are provided in the “Literacy Measure Comparability” section below.

Navy Assessment System Literacy Measure

The Navy assessment system provides an on-demand, web-based suite of diagnostic assessments designed to give reliable diagnoses of competency of the State’s challenging academic standards using a short, web-based assessment for each standard that is scored immediately to provide real-time feedback to users. Navy is a through-year assessment system where participating LEAs will take multiple assessments on an LEA-determined paced. As detailed in the original application, results of these Navy assessments will be used to produce an annual summative achievement score. Results of these assessments will also be used to produce an annual summative literacy measure. Thus, the information the Navy system collects on each student throughout the year will be summarized at the end of the year for accountability purposes, without requiring additional testing.

The Navy literacy measure will leverage a family of diagnostic psychometric methods to produce statistically driven classifications of students to describe their reading comprehension level to be (1) at or above grade level, or (2) below grade level (e.g., Bradshaw, 2017). In the Navy assessment system,

items measuring Reading Informational and Reading Literary ELA standards are based on passages that students read during the assessments. The items form the basis of the Navy's competency diagnoses for each standard; the annual summative achievement score is based upon a student's profile of competency diagnoses across the standards. These items will also form the basis of the Navy assessment system's literacy measure: Student responses to items across a sample of standards will contribute to the annual summative literacy measure.

Navy Education will examine the internal structure of the literacy measure via empirical evidence provided by the diagnostic psychometric modeling framework to assess the relationship of the item responses and measured construct. Navy Education will use the diagnostic psychometric framework to analyze results for item quality in terms of statistical information (Henson & Douglas, 2005), for differential item functioning (DIF; Hou, de la Torre, & Nandakumar, 2014; Li & Wang, 2015), and for item fit (Rupp, Templin, & Henson, 2010). Navy Education will examine the reliability of the literacy measure diagnoses and ensure the sample of items is sufficient for accountability purposes.

During the IADA period, the Putnam Consortium will include in annual technical report documentation the psychometric properties of the Navy Innovative Assessment System literacy measure.

External validity evidence of the Navy Innovative Assessment System literacy measure will be gathered via comparability studies with the Georgia Milestones results, as described in the next section.

Literacy Measure Comparability

For the literacy measure, score comparability is required at the level of the classification of (1) on-grade level or above reading comprehension or (2) below grade level reading comprehension. Comparability at this level means that evidence is provided to support the notion that if a student is determined to be an "on-grade level or above reader" in one district, had that student been assigned to another district's assessment system (for example, Navy or Georgia Milestones), he or she could expect to also be deemed an on-grade level reader. Thus, evidence is needed to support that the binary classification from Milestones' results is comparable to the binary classification from Navy's results.

The Putnam Consortium will use the same process of gathering comparability evidence for the literacy measure as the process proposed in their original application for gathering comparability evidence for the ELA annual summative determinations. The data required to examine comparability for the literacy measure is already being collected in order to examine the comparability of the ELA annual summative determinations. Namely, for comparability of annual summative determinations, the Putnam consortium's application stated that the statewide ELA assessment will be administered alongside Navy assessments in a sample of grades so that the statewide assessments can serve as a calibration tool, providing evidence about the comparability of student achievement across participating districts (giving Navy assessments) and non-participating districts.

The same concurrent and longitudinal comparability evaluations described in the Putnam consortium's application for evaluating the comparability of the annual summative determinations across assessment systems in the State will be used for evaluating the comparability of the literacy measure across assessment systems in the State. In the remainder of this section, italicized text is verbatim from the Putnam consortium's application describing concurrent and longitudinal comparability evaluations for

the annual summative determinations. New text to include also the literacy measure is given in underlined text.

*Comparability across the two assessment systems is established through **concurrent comparability evaluations**. Importantly, the degree of comparability of the annual determinations and the literacy measure classifications across the two assessment systems within the State can be directly evaluated by administering an assessment that is common across the two programs to a sample of students (i.e., by double testing) and then comparing the accuracy of proficiency and literacy measure classifications. Since the statewide academic assessment is administered once per grade span in grades 3-8 and high school, the comparability of the annual determinations and the literacy measure classifications between Navy and non-Navy districts is evaluated by directly comparing annual determinations and the literacy measure classifications for the students that participated in both assessment systems. By calculating two sets of annual determinations and two sets of literacy measure classifications for these students, the state has both traditional and innovative data points for some of the students in each Navy district. The degree of agreement between the two sets of annual determinations and two sets of literacy measure classifications is then analyzed to provide further evidence regarding the comparability of the interpretations of the reported achievement levels and the reported literacy measure classifications, or if systematic differences are detected, inform decisions about calibrating results to provide for comparability when appropriate. The degree of similarity between the proficiency and literacy classifications provides further support to the comparability of the interpretations of the reported achievement levels and literacy measure classifications across the two assessment systems. The accuracy of the proficiency and literacy classifications will be examined by grade and subject and also by waiver-reported subgroup. Results of the concurrent comparability evaluations will be reported annually to USED.*

*Comparability across the two assessment systems is established through **longitudinal comparability evaluations**. Since a sample of students participate in the statewide academic assessment once per grade span in ELA and math, we use this information to compare performance on the statewide academic assessment with performance on the Navy innovative assessments for students in certain grades and subjects where there is overlap from one year to the next. This comparison means comparing student's performance on the statewide assessment in one year to their performance in the Navy system in the next year. This also means the opposite—comparing a student's performance in the Navy system in one year to their performance on the statewide assessment in the next year. These longitudinal comparability evaluations provide evidence that the meaning of the annual determinations and literacy measure classifications is reasonably stable across years and assessment systems. We would expect the classification accuracies for the longitudinal comparability evaluations to be lower than the classification accuracies observed for the concurrent year comparisons because we would expect student achievement and student literacy levels to vary across years. Similar to the concurrent comparability evaluations, the accuracy of the proficiency and literacy classifications is examined by grade and subject and also by waiver-reported subgroup. Results of the longitudinal comparability evaluations will be reported annually to USED.*

Following suit with New Hampshire's comparability approach, results of the concurrent and longitudinal comparability analyses will be evaluated with respect to four cascading comparability questions:

1. *Are differences across the innovative and statewide assessment systems greater than differences observed across assessment conditions within the statewide assessment system?*
2. *Are the differences meaningfully significant? Do they constitute a significant threat to the validity of the accountability system or to equity in the opportunity to demonstrate knowledge of the State's challenging academic standards?*
3. *Do differences vary across subgroups or institutions in a meaningfully significant way, as to disadvantage certain subgroups or institutions?*
4. *Is the disadvantage consequential enough to not be off set by positive consequences of the innovative assessment system (e.g., positive impact on teaching and learning)?*

We will evaluate the responses to all four of these questions to consider the degree to which the assessment systems can be considered comparable enough to support their intended uses. The Putnam Consortium Innovative Assessment Executive Team in collaboration with Navy Education will take additional steps to improve the comparability of the annual summative determinations and literacy measure classifications to support the use of the innovative assessment system during the IADA period. These steps may include adjusting the performance standards (criteria or cut-offs for classifying students into Achievement Levels or into literacy measure levels) to produce comparable results for the duration of the demonstration period.

References

- Bradshaw, L. (2017). Diagnostic classification models. In D. Bandalos (Ed.), *Measurement Theory and Application for the Social Sciences*. New York, NY: Guilford Press.
- Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement, 29*(4), 262-277.
- Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald Test to investigate DIF in the DINA model. *Journal of Educational Measurement, 51*(1).
- Li, X., & Wang, W. C. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement, 52*(1), 28-54.
- Rupp, A. A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford.