New Meridian Comparability Review Guidelines

Version 6.17.2020

New Meridian Corporation has developed the *Quality Testing Standards and Criteria for Comparability Claims* (QTS). The goal of the QTS is to provide guidance to states that are interested in including content from the New Meridian item bank and intend to make comparability claims with *other assessments* that include New Meridian content. Note that "other assessments" could mean New Meridian test forms administered in another state, or it could mean the state's previous assessments that include New Meridian assessment content but administered or scored by a different testing vendor going forward. (See the Potential Use Cases section for additional information).

Comparability Review Process

One of the key assumptions is that states licensing New Meridian content are interested in comparing their assessment results to those of another assessment program. For example, to maintain trendlines, a state may wish to continue reporting scores on the New Meridian Scale in its newly developed assessment that include items from the New Meridian Item Bank. This is referred to as *scale score comparability*. Or, a state may be interested in comparing the percentage of student who are on track or ready for college and careers (attain Level 4 or higher) with other states, districts or schools that administer New Meridian content and use the New Meridian Level 4 cut score to determined college and career readiness (CCR)[1]. This is called *readiness comparability*. To support its intended comparability claims and provide for processes such as federal peer review, we recommend each interested state collect and submit evidence demonstrating that these types of comparisons are technically defensible. This evidence will be evaluated by independent expert reviewers to determine if the desired comparisons can be supported. If the desired comparisons cannot be supported, the reviewers will provide constructive and actionable feedback on what the state needs to do to support the comparability claims. This is referred to as the *QTS comparability review process*.

While the expectation is that the reviewers for this process are experts with solid technical knowledge and ample operational assessment experience, the evaluation is ultimately a judgment task. The purpose of this document, therefore, is to provide objective and concrete guidelines for experts involved in the QTS comparability review process. The overarching comparability questions that the expert reviewer is seeking to answer through his or her evaluation are:

[1] "College and career readiness" (CCR) in this context refers to both being *on track* for college and careers (i.e., grade-level readiness in elementary and middle school) and *ready* for college and careers (in high school).

- For ***scale score comparability***[2]: *If a student taking the state's assessment with New Meridian content took one of the test forms offered by New Meridian, would he or she obtain the same scale score?*
- For ***readiness comparability***[3]*: If a student taking the state's assessment with New Meridian content took one of the test forms offered by New Meridian, would he or she receive the same designation in terms of college and career readiness?*

Areas and Aspects of Evaluation

To answer these questions, the review guidelines focus on the degree to which the participating state's assessment program is comparable to the standard processes for the New Meridian test forms in three main *areas*:

- **Design**: The design of the state's assessments with New Meridian content (e.g., purpose, content representation, item types) and the procedures informing its development are comparable to those of the New Meridian test forms.
- **Administration**: The state's assessments with New Meridian content are administered under comparable conditions (with respect to factors such as testing time, directions, accommodations allowed, etc.) to those of the New Meridian test forms.
- **Scoring**: The state's assessments with New Meridian content are scored using procedures comparable to those used to score the New Meridian test forms.

Across the four areas of evaluation, the supporting evidence submitted by a state is organized and compared to various criteria in six key *aspects* of its testing program:

1. Item and Test Development (Design area)
2. Fairness and Accessibility (Design area)
3. Test Administration (Administration area)
4. Item Scoring (Scoring area)
5. Psychometrics (Scoring area)
6. Standard Setting (Scoring area)

[2] A more technical way to frame the scale score comparability question is: *Does the evidence submitted by the state for its assessments with New Meridian content meet the requirement of score equating between its forms and the New Meridian test forms*?

[3] A more technical way to frame the readiness comparability question is: *Does the evidence submitted by the state for its assessments with New Meridian content support an equivalent degree of rigor and interpretation in terms of college and career readiness as the New Meridian test forms*?

Organization of Comparability Review Guidelines

To help standardize the comparability review process, these guidelines organize the evaluation of each area into seven main sections.

- The first section is the *Bottom Line Question*, which is the overarching question that the reviewer is seeking to address through his or her evaluation of the state's evidence for this area.
- The second section is the *Key Review Consideration*, which highlights the essence of what the reviewer should focus his or her evaluation on in the review process for the two types of comparability claims.
- The third section is the *Evaluative Criteria*, which correspond to the various criteria and degree of similarity expected between the state's assessments and New Meridian test forms, as described in the QTS for each aspect of the testing program.
- The fourth section, *Sources of Evidence*, provides a list that corresponds to the examples of supporting evidence in the QTS for each aspect of the given area. This list is also reflected in the *New Meridian Comparability Evaluation Checklist*, which interested states can use to organize the documents and materials they are submitting as evidence for the comparability review process. Note that the list is not meant to be exhaustive, and not every example of supporting evidence is applicable to a given state.
- The fifth section, *Evaluative Statements*, represents the crux of the evaluation. In this section, the reviewer makes evidence-based judgments on a series of comparability statements associated with the various criteria for each aspect of the state's testing program, culminating in an overall comparability evaluation of each criterion.
- In the sixth section, *Summary*, a table is given for the reviewer to summarize his or her evaluations across the various criteria and provide suggestions for additional evidence that the state may consider submitting to strengthen its case for making comparability claims.
- In the seventh and final section, *Overall Evaluation*, the reviewer provides a response to the *Bottom Line Question* for the area, followed by feedback on the types of comparisons the state can support based on its submitted evidence, and additional comments or feedback related to the response and recommendation for the area.

QTS Documents Roadmap

This document is part of a set of supporting materials for the QTS[4]. The primary intended audience/user of this document is expert reviewers involved in the comparability review process. A roadmap of how the QTS and its supporting documents can be used in the comparability review process is illustrated in Figure 1. In this figure, the first (red) box are documents that provide information about the *standards and criteria* by

[4] Additional QTS supporting materials include the *Quality Testing Standards and Criteria for Comparability Claims* (QTS main document), *State QTS Starter Kit*, and the *Standard Processes* document.

which the state's comparability claims are evaluated. The second (green) box is the State *QTS Starter Kit*, which helps states understand the types of *evidence* that help support comparability claims. The third (purple) box is the *New Meridian Comparability Review Guidelines*, which provide concrete *guidance* on how an expert reviewer compares the evidence in the second box with the standards and criteria in the first box.

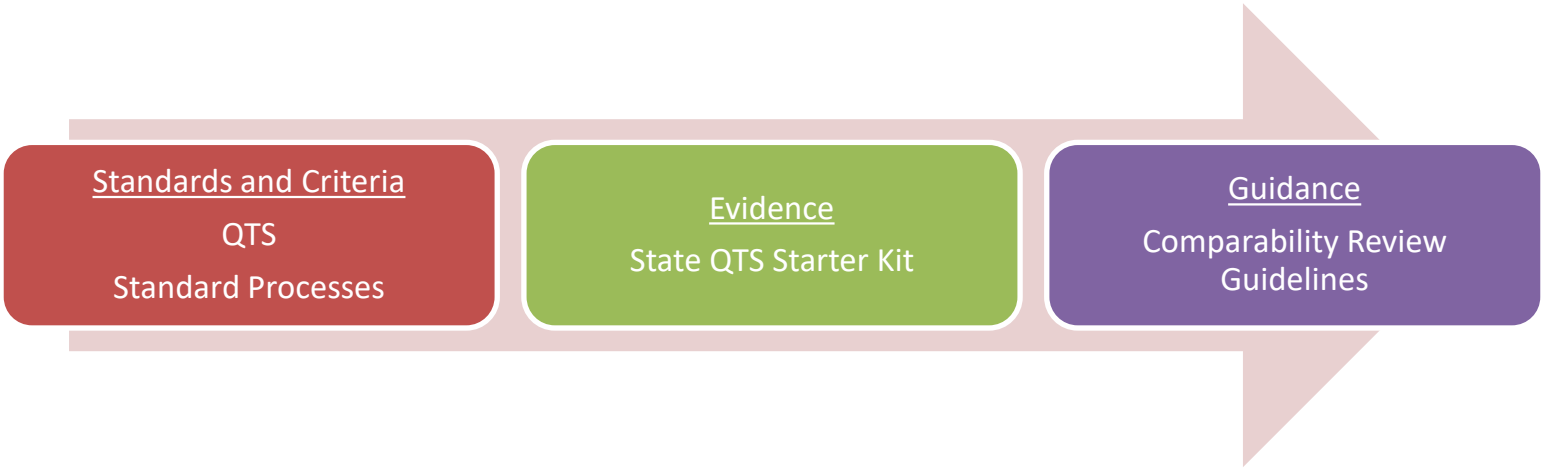

Figure 1: Suggested Roadmap for Using the QTS and its Supporting Documents

Potential Use Cases

As previously mentioned, the goal of the QTS is to provide guidance to states that are interested in including New Meridian content and intend to make comparability claims with *other assessments* that include New Meridian content. The term "other assessments" could mean something different for each state, depending on the goal for including New Meridian tasks or items on its operational test forms and the requirements for its testing program. Table 1 describes several potential use cases for tests that include New Meridian content.

Table 1: Potential Use Cases for the QTS Comparability Review Process

Use Case	Description of Use Case
State-licensed "New Meridian" forms	The state licenses New Meridian content with test forms designed to match the specifications and blueprints for New Meridian test forms. The state contracts its own vendor for the other steps in the operational administration process, including delivery and scoring.

Use Case	Description of Use Case
State-licensed "New Meridian" forms, supplemented with state-developed content	The state licenses New Meridian content, but also includes content from its own (state-specific) item bank. The test forms are designed to match the specifications and blueprints for the New Meridian test forms. The state contracts its own vendor for the other steps in the operational administration process, including delivery and scoring.
State developed assessments, supplemented with New Meridian content	The state develops its own test items but also licenses New Meridian content. The test forms are designed to match state-developed test specifications and blueprints. The state contracts its own vendor for test development, administration, and scoring.

The core question of interest and focus of the comparability evaluation differs for each of the potential use cases. It is therefore important for the expert reviews to recognize under which use case the state that they are reviewing falls. Table 2 provides the core question and focus of evaluation for each of the use cases described in Table 1.

Table 2: Core Question and Focus of Evaluation for Potential Use Cases

Use Case	Core Questions	Focus of Evaluation
State-licensed "New Meridian" forms	Are the procedures, materials, and tools used in the administration, scoring and reporting of the state-licensed "New Meridian" forms sufficiently similar to those used by the New Meridian test forms to support the use of the New Meridian scale and/or Level 4 cut score as if they were equivalent?	• Quality of adherence to the New Meridian test specifications and blueprints • Comparability in rigor and quality of procedures used to present, administer, score, and validate the assessment outcomes • Potential sources of construct irrelevant variance that would threaten the comparability of score interpretations and claims between the state's assessments with New Meridian content and the New Meridian test forms.
State-licensed "New Meridian" forms, supplemented with state-developed content	Is the construct defined by the test specifications and blueprints, the procedures used to develop and validate content, AND procedures and	Same as for the State-licensed "New Meridian" forms, with the additional key consideration of being able to support claims that the state

Use Case	Core Questions	Focus of Evaluation
	materials for administering, scoring and reporting of the state-licensed "New Meridian" forms sufficiently similar to those used by the New Meridian test forms to support the use of the New Meridian scale and/or Level 4 cut score as if they were equivalent?	developed content measures the Common Core State Standards in the same way as demonstrated on the New Meridian test forms.
State developed assessments, supplemented with New Meridian content	Is the construct defined by the test blueprint, the procedures used to develop and validate content, AND procedures and materials utilized for administering and scoring New Meridian content and reporting test results similar enough to those used by the New Meridian test forms to support the use of the New Meridian scale and/or Level 4 cut score as if they were equivalent?	Same as the State-licensed "New Meridian" forms, supplemented with state-developed content. The one key difference is rather than evaluating the quality of adherence to the New Meridian test specifications and blueprints, a focus of evaluation should be on whether the construct assessed by the state developed assessment is essentially the same as that measured by the New Meridian test forms, even though the blueprints are not the same.

If you have any questions, please contact info@newmeridiancorp.org at New Meridian Corporation.

Design

Bottom Line Question

Are the specifications and procedures underlying the design and development of the state's assessment with New Meridian content comparable to those of the New Meridian test forms?

Key Review Considerations

Scale score comparability: To support claims that scale scores resulting from the two assessments can be used interchangeably, the assessments must be purposely developed to be as similar as possible to each other in content and statistical specifications. If the assessments were developed to measure different content (e.g., standards), or measure similar content in a different way (e.g., using different item formats) or to a different extent (i.e., the relative emphasis has changed) scale scores should not be interpreted or used as if they are interchangeable even if a common subset of items exists to link the assessments to a common scale.

Readiness comparability: To support claims that specified cut scores or levels of performance on two assessments support comparable readiness inferences (i.e., on-track or ready for college and careers), the assessments should be designed to measure a common domain and the benchmark should reflect similar expectations for performance across assessments from a content perspective.

Evaluative Criteria

1. *Item and Test Development*
    a. Test purpose, target population and intended uses;
    b. Assessed content standards, item types, rubrics, blueprints, test formats, eligible content, and time limits, along with the rationale for the test design decisions;
    c. Procedures for review of test items by subject matter experts;
    d. Field testing and data review procedures; and,
    e. Forms construction and review procedures.
2. *Fairness and Accessibility*
    a. Universal design principles;
    b. Accommodations for English learners and students with disabilities; and,
    c. Procedures used to translate forms for students for whom English is a second language.

Table 3 below is taken from the QTS and summarizes the degree of similarity expected between a state's test design and New Meridian's for each type of comparability claims. The table also indicates whether the state's supporting evidence for each criterion should be considered in the evaluation of the two types of comparability claims. "Yes" means that evidence of comparability between the assessments for this criterion is required; while "Recommended" means that evidence of comparability for this criterion would help the state's case. Even if the cell for a criterion is blank, the reviewer is encouraged to consider the evidence submitted by the state as it may be informative, either in support of or against, the state's case for making comparability claims.

Table 3: Criteria for Making Comparability Claims – Design Area

Criterion/ Type of Comparability Claim	Scale Score Comparability	Readiness Comparability
Overall Degree of Similarity with New Meridian Design	**High**	**Adequate**
Criterion 1a Test purpose and uses	Recommended	Recommended
Criterion 1b Test and item specifications	Yes	Yes
Criterion 1c Evaluation of new items	Recommended	
Criterion 1d Forms construction	Yes	Yes
Criterion 2a Universal design	Recommended	Recommended
Criterion 2b Accommodations and language translation	Recommended	Recommended

Sources of Evidence

Below is a list of potential evidence that the state may provide to support its comparability to the New Meridian test forms in the **Design** area. Please refer to the completed *Comparability Evaluation Questionnaire* and *Comparability Evaluation Checklist* for information provided by the state about its testing program and submitted evidence.

Item and Test Development

- ✓ Documentation or web pages about the testing program and its assessments
- ✓ Documentation or web pages about the assessed curriculum (state-mandated or district-selected), content standards, and claims structure
- ✓ Item development specifications and processes, and qualitative and quantitative item review and piloting procedures
- ✓ Test development and review procedures, including test blueprints or specifications
- ✓ Forms or test construction specifications, including test construction targets, and forms review and approval procedures
- ✓ Materials or minutes for educator or stakeholder committee meetings
- ✓ Content alignment study reports

Fairness and Accessibility

- ✓ Accommodations manuals, tutorials or guides
- ✓ Test translation or transadaptation guidelines
- ✓ Materials or minutes from bias and sensitivity review committee meetings
- ✓ Evidence supporting the fairness of assessment results for all students and disaggregated student groups
- ✓ Research reports related to accessibility, universal design principles, and the validity of accommodations and language translations
- ✓ Annual technical reports or manuals

Evaluative Statements

Each criterion includes superscripted symbols that indicates whether evidence of comparability between the state's assessment and the New Meridian test forms is required ("Yes" in Table 3) or recommended ("Recommended' in Table 3). An upper-case "S" or "R" means that evidence of comparability on this criterion is required for scale score or readiness comparability respectively; while a lower-case "s" or "r" means that evidence of comparability on this criterion is recommended for scale score or readiness comparability, respectively.

Item and Test Development

1a. Test Purpose and Uses [sr]

✓ The state assessment program is designed to make similar claims about student achievement as the New Meridian test forms[5].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence

Reviewer Comments:

✓ The state assessment program is intended to be taken by a similar student population (in terms of grade levels, student groups, special populations, etc.) as the New Meridian test forms[6].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence

Reviewer Comments:

✓ Results from the state assessment program are intended to be used by the state in similar ways as results are by states that administer the New Meridian-designed assessment forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence

Reviewer Comments:

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **test purpose and uses**?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar

[5] The claims (and sub-claims) of the New Meridian test forms can be found in section 2.1 of the 2019 Technical Report (Flagship, ABO).

[6] A summary of the test-taking population for the New Meridian form can be found in section 11 of the 2019 Technical Report (Flagship, ABO).

Reviewer Comments:

1b. Test and Item Specifications [SR]

✓ The state assessment is designed to assess a highly similar set of content standards to those assessed by the New Meridian test forms.[7]

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence

Reviewer Comments:

✓ The blueprints for the state assessments are highly similar to the blueprints[8] for the New Meridian test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence

Reviewer Comments:

✓ The state assessment includes similar item types as those on the New Meridian test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence

Reviewer Comments:

[7] Evidence statement, which describe the knowledge and skills assessed at each content area and grade level of the New Meridian test forms, are available on at the following pages for ELA/Literacy (scroll down to the *Reading Evidence Tables* and *Writing Evidence Tables* sections) and Mathematics (scroll down to the *Evidence Statement Documents* section.)

[8] Blueprints for the New Meridian test forms are available at the following pages. For the Flagship forms: ELA/L Grades 3-5, ELA/L Grades 6-11, and Mathematics; for the ABO forms: ELA/L and Mathematics

✓ If the state is using prose constructed response (PCR) writing tasks from the New Meridian bank on its state assessment, then it is using the same scoring rubric[9] as the one used for PCRs on the New Meridian ELA/Literacy forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ The presentation of the content (including format, layout, style, etc.) on test forms for the state assessments are not substantially different to that on the New Meridian test forms such that it changes the underlying nature of the constructed being measured.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence

Reviewer Comments:

✓ If the state assessment is administered as an adaptive assessment – that is, as a multistage test (MST) or item-level computer adaptive test (CAT) – the state has evidence (e.g., adequacy of its item pool to represent the test specifications and blueprints, adequacy of the adaptive algorithm to produce an aligned, representative test forms across the full ability distribution, etc.) that the design and algorithm of the MST or CAT provide for "test forms" that represent the same content domains and construct as assessed by the New Meridian form.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Does Not Apply

Reviewer Comments:

✓ The testing times and testing window for the state assessments are similar to that of the New Meridian test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

[9] The scoring rubrics for PCR items can be found on the ELA/Literacy page (scroll down to the *ELA/Literacy Scoring Rubrics* section.)

Reviewer Comments:

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **test and item specifications**?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar

Reviewer Comments:

1c. Evaluation of New Items [s]

✓ Newly developed items on the state assessment go through a similar review process as the New Meridian test forms, which include the following committees of subject matter experts prior to field-testing.

- Text Review Committee
- Content Review Committee
- Bias and Sensitivity Item Committee
- Editorial Review Committee

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ The approach and procedures used to field-test newly developed items on the state assessment provide for performance data that is at a comparable level of quality as that collected for the New Meridian summative assessments (e.g., a large representative sample of students take each question under operational conditions).[10]

[10] More details about the New Meridian field test data collection approach can be found in section 2.2.5 of the 2019 Technical Report (Flagship, ABO).

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ The state assessment program implements an item-level data review process that is comparable to that implemented for New Meridian items[11] in terms of item statistics included, criteria for accepting/rejecting items, and data review committee composition.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **evaluation of new items**?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar

Reviewer Comments:

1d. Forms Construction SR

✓ The state assessment program involves subject matter experts and educators in the construction, review and approval of its operational test forms in a similar manner as that for the New Meridian test forms[12].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

[11] Information about the New Meridian data review process is available in section 2.2.2 of the 2019 Technical Report (Flagship, ABO).

[12] Details of content expert and educator involvement in the forms construction process for the New Meridian test forms are available in section 2.2.3 of the 2019 Technical Report (Flagship, ABO).

✓ The state assessment program uses similar psychometric criteria and/or targets for the construction and review of its operational test forms as the New Meridian test forms[13]. More specifically, the key psychometric criteria used in the construction process for the New Meridian test forms include:

- Evaluation of test characteristic curves (TCCs) across the score range and at each cut score against predetermined targets;
- Evaluation of test information function (TIF) curves across the score range and at each cut score against predetermined targets;
- Evaluation of conditional standard error of measurement (CSEM) curves across the score range and at each cut score against predetermined targets;
- Examination of classical item statistics, including the distribution of average item scores (or p-values) and item-total correlations;
- Examination of IRT statistics, including the distribution of discrimination (a) and difficulty (b) parameters;
- Review of items flagged for differential item functioning (DIF).

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ The state assessment program uses a similar process for building accommodated operational test forms as the New Meridian test forms[14].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

[13] For readiness comparability, the evaluation only needs to be on the key psychometric criteria as they related to the college and career readiness benchmark.
[14] Information about New Meridian's accommodated forms review process is in section 2.2.3 (*Accommodated Forms Review Process* subsection) of the 2019 Technical Report (Flagship, ABO).

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **forms construction**?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar

Reviewer Comments:

Fairness and Accessibility

2a. Universal Design [sr]

✓ The state assessment program follows universal design requirements that are comparable to those adhered to in the item and operational form development process for the New Meridian test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence

Reviewer Comments:

✓ The state assessment program provides similar tools and accessibility features to all students as those on the New Meridian test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence

Reviewer Comments:

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **universal design**?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar

Reviewer Comments:

2b. Accommodations and Language Translation [sb]

✓ The state assessment program has similar participation policies and guidelines for accommodations as those defined for the New Meridian test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ The state assessment program offers a similar set of accommodations (online and paper-based) for students with disabilities (SWD) as those available on the New Meridian test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ The state assessment program offers a similar set of accommodations (online and paper-based) for English learners (EL) as those available on the New Meridian test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ If the state assessments with New Meridian content are offered in other languages, the state has processes in place to validate the accuracy of language translation for its operational test forms[15].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **accommodations and language translation**?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar

Reviewer Comments:

Summary

In Table 4, please summarize the comparability ratings you gave for each criterion in the previous section and indicate your level of confidence in each rating based on the evidence submitted by the state for the **Design** area. If you are not very confident about a rating, please suggest additional evidence that the state could provide to support its comparability with the New Meridian test forms for the criterion.

As in the previous section, each criterion includes superscripted symbols that indicates whether evidence of comparability between the state's assessment and the New Meridian test forms is required ("Yes" in Table 3) or recommended ("Recommended' in Table 3). An upper-case "S" or "R" means that evidence of comparability on this criterion is required for scale score or readiness comparability respectively; while a lower-case "s" or "r" means that evidence of comparability on this criterion is required for scale score or readiness comparability, respectively.

Table 4: Summary of Comparability Ratings for the Design Area

[15] Information about the translation (or transadaptation) process for Spanish New Meridian test forms is available in section 2.2.3 (*Spanish-Language Assessments for Mathematics* subsection) of the 2019 Technical Report (Flagship, ABO).

Criterion	Comparability rating	Confidence in rating	What additional evidence could be provided? (If somewhat or not confident in your rating)
1a – Test purpose and uses [sr]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	
1b – Test and item specifications [SR]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	
1c – Evaluation of new items [s]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	
1d – Forms construction [SR]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	
2a – Universal design [sr]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	
2b – Accommodations and language translation [sr]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	

Overall Evaluation

Based on your review of the submitted evidence, what is your response to the *Bottom Line Question* for the **Design** area:

*Overall, do you believe that the specifications and procedures underlying the design and development of the state's assessment with New Meridian content is similar enough (i.e., **adequately similar**) to those of the New Meridian test forms to support **Readiness Comparability**?*

☐ Yes ☐ No ☐ More Information Needed

*Overall, do you believe that the specifications and procedures underlying the design and development of the state's assessment with New Meridian content is similar enough (i.e., **highly similar**) to those of the New Meridian test forms to support **Scale Score Comparability**?*

☐ Yes ☐ No ☐ More Information Needed

Please provide any additional comments or feedback to support your overall comparability evaluation of the **Design** area.

Administration

Bottom Line Question

Are the policies and procedures for administering the state's assessment that include New Meridian content comparable to those of the New Meridian test forms?

Key Review Considerations

For *both* **scale score comparability** and **readiness comparability**: There are a multitude of test administration factors that can call into question the comparability of assessment results, even if tests were essentially designed to adhere to the same content and statistical specifications. If administration factors differ greatly with respect to such aspects as training and instructions given to test administrators and test takers, mode of administration, testing supports (e.g., tools, accommodations and opportunities for practice) and testing time (i.e., duration and timing/length of administration windows), then scale scores and readiness benchmarks resulting from the two assessments should not be considered interchangeable.

Evaluating comparability with respect to test administration is therefore predominantly *process oriented*. When reviewing the evidence for this area, consider whether the state has established clear test administration policies and procedures, communicated those policies and procedures effectively, and put systems in place to monitor their implementation. Use the test administration policies and procedures for the New Meridian test forms as a standard against which to compare, but realize that each state may have specific requirements or constraints that necessitate differences in its test administration processes. In such cases, think about whether the differences are such that they would affect the underlying nature of the constructed being measured.

Evaluative Criteria

3. *Test Administration*
    a. Training and instructions provided to test administrators and coordinators;
    b. Instructions given to test takers;
    c. Information about the modes of administration, such as paper-based vs. computer-based testing, and fixed-form vs. adaptive tests, including rationale for the offering the test in each mode;
    d. Details about test security protocols; and,
    e. Evidence that supports accessibility of the test to all students as part of the test administration.

Table 5 below is taken from the QTS and summarizes the degree of similarity expected between a state's test administration processes and New Meridian's for each type of comparability claims. As noted in the *Key Review Considerations* section, there is no clear distinction in the Administration area on the degree of similarity required for the two types of comparability claims. As such, the values in the columns for Scale Score Comparability and Readiness Comparability are identical. As with Table 3 for the Design area, "Yes" means that evidence of comparability between the assessments for this criterion is required; while "Recommended" means that evidence of comparability for this criterion would help the state's case.

Table 5: Criteria for Making Comparability Claims – Administration Area

Criterion/ Type of Comparability Claim	Scale Score Comparability	Readiness Comparability
Overall Degree of Similarity with New Meridian Administration Processes	**High**	**High**
Criterion 3a Administrator guidance	Yes	Yes
Criterion 3b Test taker instructions	Yes	Yes
Criterion 3c Administration modes and designs	Recommended	Recommended
Criterion 3d Security protocols	Recommended	Recommended
Criterion 3e Accessibility supports	Yes	Yes

Sources of Evidence

Below is a list of potential evidence that the state may provide to support its comparability to New Meridian in the **Administration** area. Please refer to the completed *Comparability Evaluation Questionnaire* and *Comparability Evaluation Checklist* for information provided by the state about its testing program and submitted evidence.

Test Administration

- ✓ District and/or campus test administrator and coordinator manuals and training materials
- ✓ Practice tests or tutorials for test takers
- ✓ Documentation or web pages about online testing interface for computer-based tests
- ✓ Specification about technology requirements for computer-based tests
- ✓ Research reports on administration mode (paper vs. computer) or device (computer vs. tablet) comparability studies
- ✓ Specification about adaptive testing procedures or process
- ✓ Test security and administration procedures
- ✓ Documentation about data forensics analyses
- ✓ Accommodations manuals, tutorials or guides

Evaluative Statements

Each criterion includes superscripted symbols that indicates whether evidence of comparability between the state's assessment and the New Meridian test forms is required ("Yes" in Table 5) or recommended ("Recommended' in Table 5). An upper-case "S" or "R" means that evidence of comparability on this criterion is required for scale score or readiness comparability respectively; while a lower-case "s" or "r" means that evidence of comparability on this criterion is recommended for scale score or readiness comparability, respectively.

Test Administration

3a. Administrator Guidance ^{SR}

✓ The clarity and level of specificity of the instructions given to local test administrators and coordinators to plan, administer, and complete the state's assessment are highly similar to those used for the New Meridian test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence

Reviewer Comments:

✓ The quality and rigor of procedures and materials used to train local test administrators and coordinators of the state's assessment are highly similar to those for the New Meridian test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence

Reviewer Comments:

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **administrator guidance**?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar

Reviewer Comments:

3b. Test Taker Instructions [SR]

✓ The clarity and level of specificity of the verbal instructions given to the test takers by the test administrator for the state's assessment are highly similar to those presented for the New Meridian test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence

Reviewer Comments:

✓ If the state assessments include paper-based test forms, the clarity and level of specificity of the printed directions for test takers are highly similar to those on the paper-based New Meridian test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Does Not Apply

Reviewer Comments:

✓ If the state assessments include computer-based test forms, the breadth and depth of the topics covered in the tutorials for the online testing interface and various item types are highly similar to those on the computer-based New Meridian test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Does Not Apply

Reviewer Comments:

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **test taker instructions**?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar

Reviewer Comments:

3c. Administration Modes and Designs [sr]

✓ If the state assessment includes computer-based test (CBT) forms, the state has evidence (e.g., how items are rendered and behave in the online testing interface) to support the comparability of New Meridian CBT items on the state assessment and the New Meridian test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Does Not Apply

Reviewer Comments:

✓ If the state assessment includes paper-based test (PBT) forms, the state has evidence (e.g., how items appear in the printed test forms) to support the comparability of New Meridian PBT items on the state assessment and the New Meridian test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Does Not Apply

Reviewer Comments:

✓ If the state assessment is administered as both CBT and PBT, the state has evidence (such as results from empirical research studies) to support the comparability of scale scores from across the two modes[16].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Does Not Apply

Reviewer Comments:

✓ If the state assessment is administered as an adaptive assessment – that is, as a multistage test (MST) or item-level computer adaptive test (CAT) – the state has evidence that the administration rules and policies that are in place to accommodate the use of an adaptive engine (e.g., no review of items, no changing previous responses, etc.) are implemented in a way that does not influence the comparability of the students responses with those collected from the non-adaptive test form, such as those on the New Meridian test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Does Not Apply

Reviewer Comments:

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **administration modes and designs**?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar

Reviewer Comments:

[16] As an example, mode comparability studies have been conducted for the CBT and PBT New Meridian test forms. A high-level description of the study design and findings is available in section 14.5.4 of the 2019 Technical Report (Flagship, ABO).

3d. Security Protocols [sr]

✓ The state assessment includes policies and procedures for maintaining the security of test materials and content that are as rigorous and comprehensive as those for the New Meridian test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence

Reviewer Comments:

✓ The state assessment includes protocols for handling testing irregularities and security breaches are as rigorous and comprehensive as those for the New Meridian test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Does Not Apply

Reviewer Comments:

✓ The state assessment conducts empirical analyses of student response to detect possible testing irregularities (i.e., data forensics analyses) that are as rigorous and comprehensive as those conducted for the New Meridian test forms[17].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Does Not Apply

Reviewer Comments:

✓ If the state assessment is administered as an adaptive assessment – that is, as a multistage test (MST) or item-level computer adaptive test (CAT) – the state has appropriate exposure control procedures in place to ensure items are not overexposed to test takers and test materials are kept as secure.

[17] Overviews of the data forensics analysis methods for the New Meridian test forms are in section 3.4 of the 2019 Technical Report (Flagship, ABO). The analysis methods include response change analysis, plagiarism analysis, internet and social media monitoring, and off-hours testing monitoring.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Does Not Apply

Reviewer Comments:

Given the preponderance of evidence, how would you rate the degree of rigor reflected in the **security protocols** used by the state and required by New Meridian?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar

Reviewer Comments:

3e. Accessibility Supports SR

✓ If the state assessment includes CBT forms, the guidelines, directions and eligibility requirements for administering online accessibility features and accommodations for the state assessment are similar in rigor and comprehensiveness to those for the New Meridian CBT forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence

Reviewer Comments:

✓ If the state assessment includes PBT forms, the guidelines, directions and eligibility requirements for administering accessibility features and accommodations for the state's PBT forms are similar in rigor and comprehensiveness to those for the New Meridian PBT forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Does Not Apply

Reviewer Comments:

✓ The state assessment program provides resources to help students, teachers and parents become familiar with the tools and accessibility features on its test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence

Reviewer Comments:

✓ The state assessment program provides resources to help SWD and EL become familiar with the accommodations on its test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence

Reviewer Comments:

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **accessibility supports**?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar

Reviewer Comments:

Summary

In Table 6 below, please summarize the comparability ratings you gave for each criterion in the previous section and indicate your level of confidence in each rating based on the evidence submitted by the state for the **Administration** area. If you are not very confident about a rating, please suggest additional evidence that the state could provide to support its comparability with the New Meridian test forms for the criterion.

As in the previous section, each criterion includes superscripted symbols that indicates whether evidence of comparability between the state's assessment and the New Meridian test forms is required ("Yes" in Table 5) or recommended ("Recommended' in Table 5). An upper-case "S" or

"R" means that evidence of comparability on this criterion is required for scale score or readiness comparability respectively; while a lower-case "s" or "r" means that evidence of comparability on this criterion is required for scale score or readiness comparability, respectively.

Table 6: Summary of Comparability Ratings for the Administration Area

Criterion	Comparability rating	Confidence in rating	What additional evidence could be provided? (If somewhat or not confident in your rating)
3a – Administrator guidance [SR]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	
3b – Test taker instructions [SR]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	
3c – Administration modes and designs [sr]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	
3d – Security protocol [sr]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	
3e – Accessibility supports [SR]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	

Overall Evaluation

Based on your review of the submitted evidence, what is your response to the *Bottom Line Question* for the **Administration** area:

*Overall, do you believe that the policies and procedures for administering the state's assessment with New Meridian content are **highly similar** to those of the New Meridian test forms to support both **Scale Score Comparability** and **Readiness Comparability**?*

☐ Yes ☐ No ☐ More Information Needed

Please provide any additional comments or feedback to support your overall comparability evaluation of the **Administration** area.

Scoring

Bottom Line Question

Are the criteria, methodologies and procedures for scoring the state's assessment with New Meridian content comparable to those of the New Meridian test forms?

Key Review Considerations

Scale score comparability: To support claims that scaled scores resulting from the two assessments can be used interchangeably, the criteria, methodologies and procedures for scoring the assessments must be purposely developed and implemented to be as similar as possible to each other. This includes the processes for scoring item responses and for assigning scaled scores (i.e., scaling and equating).

When reviewing the evidence in this area for scale score comparability, consider first whether the state's rules and procedures for scoring items, including machine- and human-scored items, could yield systematically different scores. Then evaluate whether the state's underlying assumptions, methodological choices and specifications for scaling and equating could lead to score scales that are characteristically different. In such cases, scale scores should not be interpreted or used as if they are interchangeable, even if a common subset of items is included to link the state's assessment to the scale of the New Meridian test forms.

Readiness comparability: To support claims that specified cut scores or levels of performance on two assessments support comparable readiness inferences (i.e., on-track or ready for college and careers), the definitions of readiness and the process for establishing performance levels based on the definitions should be similar across the assessments.

When reviewing the evidence in this area for readiness comparability, first compare the policy definitions and performance level descriptors (PLDs) for readiness (i.e., on-track or ready for college and careers) between the state's assessment and New Meridian test forms. Then evaluate whether the state has provided sufficient validity evidence that the readiness benchmark on its assessment are of equivalent rigor and can be interpreted in a similar manner as the college and career readiness benchmark (Level 4) on the New Meridian test forms.

Evaluative Criteria

4. *Item Scoring*
    a. Training and qualification procedures for human scorers;
    b. Protocols for both machine and human scoring processes;
    c. Evidence that the scoring process is fair to all students; and,

   d. If used, validation of automated scoring processes.

5. *Psychometrics*
   a. Choice of psychometric models;
   b. Scaling and equating design and procedures, including quality control processes;
   c. Analysis of disaggregated student groups;
   d. Sampling, including purpose and methodology; and,
   e. Other psychometric procedures or analyses that support the reliability and validity of test scores.
6. *Standard Setting*[18]
   a. Achievement or performance level descriptors (ALDs or PLDs) and how they were established;
   b. Standard setting methodology and procedures; and,
   c. Empirical support for the readiness (on-track or ready for college and careers) cut scores.

Table 7 below is taken from the QTS and summarizes the degree of similarity expected between a state's scoring processes and New Meridian's for each type of comparability claims. The table also indicates whether the state's supporting evidence for each criterion should be considered in the evaluation of the two types of comparability claims. "Yes" means that evidence of comparability between the assessments for this criterion is required; while "Recommended" means that evidence of comparability for this criterion would help the state's case. Even if the cell for a criterion is blank, the reviewer is encouraged to review the evidence submitted by the state as it may be informative, either in support of or against, the state's case for making comparability claims.

Table 7: Criteria for Making Comparability Claims – Scoring Area

Criterion/ Type of Comparability Claim	Scale Score Comparability	Readiness Comparability
Degree of Similarity with New Meridian Scoring Processes	**High**	**Adequate**
Criterion 4a Scorer qualification	Yes, if applicable	

[18] In general, the criteria for *Standard Setting* applies only to the evaluation of readiness comparability. The rationale is that if a state has provided evidence to support scale score comparability, then it can make use of the existing New Meridian cut scores, which were established on the New Meridian scale. If the state, however, makes any adjustments to the existing New Meridian performance standards (e.g., modifies PLDs, changes cut scores etc.), then a standards validation process would likely be needed. Evidence from such a process should be provided for review based on criteria 6a to 6c.

Criterion/ Type of Comparability Claim	Scale Score Comparability	Readiness Comparability
Criterion 4b Scoring protocols	Yes	
Criterion 4c Fairness in scoring	Recommended	
Criterion 4d Automated scoring	Yes, if applicable	
Criterion 5a Psychometric models	Yes	
Criterion 5b Scaling and equating	Yes	
Criterion 5c Student group analysis	Recommended	
Criterion 6a PLDs	Recommended, if applicable[19]	Yes
Criterion 6b Standard setting method	Recommended, if applicable[20]	Yes
Criterion 6c Validity of cut scores	Recommended, if applicable[21]	Yes

Sources of Evidence

Below is a list of potential evidence that the state may provide to support its comparability to the New Meridian form in the **Scoring** area. Please refer to the completed *Comparability Evaluation Questionnaire* and *Comparability Evaluation Checklist* for information provided by the state about its testing program and submitted evidence.

Item Scoring

- ✓ Documentation about machine scoring rules, test maps, test deck, and quality assurance procedures
- ✓ Documentation about recruitment and qualification of human scorers
- ✓ Training materials for human scorers

[19] See previous footnote.

- ✓ Procedures for calibrating scoring throughout the human scoring process
- ✓ Procedures and criteria for monitoring human scorer quality
- ✓ Sample scoring materials, including rubric and anchor, training, qualifying, and validity sets
- ✓ Reports about the human scoring process, including inter-rater reliability, score point distribution, and validity sets results
- ✓ Research reports about the validity of automated (AI) scoring and fairness of the scoring process to all students
- ✓ Annual technical reports or manuals

Psychometrics

- ✓ Operational psychometrics procedures specifications or guidelines
- ✓ Specification about adaptive testing methodology (routing logic, stopping rules, content balancing and exposure control criteria, etc.)
- ✓ Equating and scaling specifications, including quality assurance procedures and criteria
- ✓ Documentation about the choice of measurement model, how scales were established, and scale score characteristics
- ✓ Documentation about sampling for scaling, equating, or other psychometric analyses
- ✓ Procedures and results of any analysis of disaggregated student group performance on operational items
- ✓ Analysis or studies that support the reliability and validity of test scores
- ✓ Research plans or reports that support the comparability of test scores between the state's assessment and New Meridian
- ✓ Annual technical reports or manuals

Standard Setting

- ✓ Procedures for establishing policy descriptors, and ALDs or PLDs
- ✓ Research studies or documentation that supports the standard setting methodology and procedures
- ✓ Standard setting specifications that include details about each step of the standard setting process
- ✓ Description of all stakeholders involved in the standard setting process
- ✓ Standard setting meeting materials, including agenda, facilitator slides, panelist forms, and example of feedback data
- ✓ Studies or empirical data that support the validity of cut scores across grade levels and/or content areas
- ✓ External validity research studies, such as correlational, linking and benchmarking studies
- ✓ Research reports on consequential validity
- ✓ Standard setting technical report or summary

Evaluative Statements

Each criterion includes superscripted symbols that indicates whether evidence of comparability between the state's assessment and the New Meridian test forms is required ("Yes" in Table 7) or recommended ("Recommended' in Table 7). An upper-case "S" or "R" means that evidence of comparability on this criterion is required for scale score or readiness comparability respectively; while a lower-case "s" or "r" means that evidence of comparability on this criterion is recommended for scale score or readiness comparability, respectively.

Item Scoring

4a. Scorer Qualification [S]

✓ The quality and rigor of procedures and materials used to train human scorers for the state's assessment are similar to those for the New Meridian test forms[20].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ The state's criteria for qualifying human scorers of ELA/literacy writing items are similar in rigor to those used by New Meridian for its operational test forms[21].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ The state's criteria for qualifying human scorers of mathematics constructed responses items are similar in rigor to those used by New Meridian for its operational test forms[22].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

[20] The process for developing scorer training materials is described in section 4.2.1 of the 2019 Technical Report (Flagship, ABO).
[21] The scorer qualification process and criteria for the ELA/literacy writing items are described in section 4.2.2 of the 2019 Technical Report (Flagship, ABO).
[22] The scorer qualification process and criteria for the mathematics constructed response items are described in section 4.2.2 of the 2019 Technical Report (Flagship, ABO).

Reviewer Comments:

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **scorer qualification**?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar ☐ Does not apply

Reviewer Comments:

4b. Scoring Protocols[S]

✓ For machine-scored items, the state assessment includes procedures for validating the answer keys and scoring rules that are similar to those for the New Meridian test forms[23].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ If the state used the New Meridian-supplied test deck[24] to validate its machine-scoring process, the resulting scores for the item responses in test deck are the same as those yielded by the machine-scoring process for the New Meridian test forms.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

[23] Descriptions of the review process for key based and rule based machine-scored items are in section 4.1 of the 2019 Technical Report (Flagship, ABO).

[24] The test deck contains simulated responses to items in each content area and language (English or Spanish) version of the mathematics test. The test deck contains examples of all relevant responses to items with complex scoring rules, unusual response patterns (e.g., multiple responses to a single item, blanks), variations on responses to gridded response mathematics items, as well as partially completed and partially erased responses (for paper-based responses). The goal of the test deck is to validate that the scoring rules and scanning rules and procedures for machine-scored items are being applied correctly.

Reviewer Comments:

✓ The state assessment implements double scoring (i.e., two human scorers, or one human + one AI scorer) for student responses to the New Meridian prose-constructed response (PCR) items. The rules and procedures for resolving scorer disagreement are similar to those for the New Meridian test forms[25].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ The state assessment implements procedures for monitoring scorer quality that are highly similar to those for the New Meridian test forms[26].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ For items from the New Meridian bank, the state's scoring protocols for both machine- and human-scores items yield the same scores as those from the New Meridian scoring process.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

[25] The rules and procedures for the double scoring of PCR items are provided in section 4.2.4 of the 2019 Technical Report (Flagship, ABO).

[26] The procedures for monitoring human scoring, including backreading, validity sets, and the use of calibration responses and inter-rater agreement statistics for scorer intervention are described in section 4.2.4 of the 2019 Technical Report (Flagship, ABO).

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **scoring protocols**?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar ☐ Does not apply

Reviewer Comments:

4c. Fairness in Scoring [S]

✓ The state assessment has procedures and criteria in place to evaluate the quality and comparability of performance scoring procedures across demographic student groups that are similar to those used by New Meridian for its operational test forms[27].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **fairness in scoring**?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar ☐ Does not apply

Reviewer Comments:

4d. Automated Scoring [S]

✓ The state assessment uses an automated scoring product or engine that has been shown to generate valid and reliable scores.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

[27] A description of the analysis to compare scoring quality for various demographic student groups is given in the 2019 Technical Report (Flagship, ABO).

Reviewer Comments:

✓ For items from the New Meridian bank, the state's automated scoring product or engine provides the same scores as expected.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ The way in which the state assessment uses scores from the automated scoring engine to assign the final score for a student response is similar to that for the New Meridian test forms[28].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **automated scoring**?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar ☐ Does not apply

Reviewer Comments:

[28] Rules for assigning final scores to PCR responses that are AI-scores are outlined in the 2019 Technical Report (Flagship, ABO).

Psychometrics

5a. Psychometric Models[S]

✓ The state assessment uses the two-parameter logistic (2PL) item response theory (IRT) model and generalized partial credit (GPC) model as its underlying measurement model (theta scale)[29].

☐ Agree ☐ Disagree ☐ Does Not Apply

Reviewer Comments:

✓ There is evidence that the IRT calibration software produces parameter estimates that are consistent (within acceptable tolerance) with those generated by IRTPRO, the IRT calibration software used for the New Meridian test forms[30].

☐ Agree ☐ Disagree ☐ Does Not Apply

Reviewer Comments:

✓ To estimate item parameters for a given administration, the state assessment uses calibration procedures and convergence criteria similar to those for the New Meridian test forms[31].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

[29] A description of the 2PL IRT and GPC models used as the underlying measurement model for the New Meridian item bank is given in section 7.3.1 of the 2019 Technical Report (Flagship, ABO).

[30] According to section 7.3.4 of the 2019 Technical Report (Flagship, ABO), IRTPRO for Windows (version 4.2) is used for the calibration process of the New Meridian bank. More information about IRTPRO is available at the Vector Psychometric Group web page.

[31] A description of the calibration procedures, including item exclusion rules, and convergence criteria for the New Meridian test forms is provided in sections 7.3.3 and 7.3.4 of the 2019 Technical Report (Flagship, ABO).

✓ The state assessment uses the same approach to estimate trait parameters for the ELA/literacy prose constructed response (PCR) item as that for the New Meridian test forms[32].

☐ Agree ☐ Disagree ☐ Does Not Apply

Reviewer Comments:

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **psychometric models**?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar ☐ Does not apply

Reviewer Comments:

5b. Scaling and Equating[S]

✓ Like the New Meridian test forms, the state assessment's scale links back to the spring 2016 online (CBT) IRT base scale.

☐ Agree ☐ Disagree ☐ Does Not Apply

Reviewer Comments

✓ If the state assessment uses a post-equating model, its method and criteria of obtaining and evaluating the post-equating samples are technically defensible and appropriate[33].

[32] A description of the approach for calibrating PCR items to account for the local dependency of traits is given in section 7.3.2 of the 2019 Technical Report (Flagship, ABO).

[33] An example of the method and criteria for obtaining early samples for post-equating is provided in section 7.2.1 of the 2019 Technical Report (Flagship, ABO).

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ The state assessment uses the same scale transformation constants to generate raw-to-scale score conversion tables as those for the New Meridian test forms[34].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ The state assessment's linking design and scaling procedure for year-to-year equating is similar to that of the New Meridian test forms[35].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ The state assessment's linking design across administration modes (CBT and PBT) is similar to that of the New Meridian test forms[36].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

[34] A description of the process for generating score conversion tables, including the scale transformation constants, for the New Meridian test forms is given in section 12.3 of the 2019 Technical Report (Flagship, ABO).

[35] The linking design for the operational New Meridian test forms is described in section 2.2.4 of the 2019 Technical Report (Flagship, ABO). The scaling procedure for year-to-year equating is described in section 7.6 of the 2019 Technical Report (Flagship, ABO).

[36] The linking design for the operational New Meridian test forms is described in section 2.2.4 of the 2019 Technical Report (Flagship, ABO).

✓ The state assessment uses quality control procedures for its calibration, scaling and equating processes that are as rigorous as those used for the New Meridian test forms[37].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ If the state has conducted or is planning to conduct any research studies that have scoring implications (for example, CBT vs. PBT or digital devices comparability studies) for its assessment with New Meridian content, its method and criteria of obtaining and evaluating the study sample are technically defensible and appropriate[38].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **scaling and equating**?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar ☐ Does not apply

Reviewer Comments:

[37] The quality control procedures for the calibration, scaling and equating of New Meridian test forms is outlined in section 7.7 of the 2019 Technical Report (Flagship, ABO).

[38] As examples, section 14.5.4 of the 2019 Technical Report (Flagship, ABO) provides overviews of the mode and device comparability studies, including how the study samples were obtained.

5c. Student group analysis [s]

✓ The state assessment's annual scaling and equating process includes analyses of disaggregated student group performance on operational items that are similar to those for the New Meridian test forms[39].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ As with the New Meridian test forms, results for the state assessment's student group analyses of operational items are reviewed by content experts to inform whether items with potential bias are excluded from the generation of score conversion tables and score reporting[40].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **student group analysis**?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar ☐ Does not apply

Reviewer Comments:

[39] A description of the analysis of disaggregated student group performance conducted annually on operational items in the New Meridian bank is given in sections 6.2 (DIF procedures and classification criteria) and 6.3 (DIF comparison groups and sample size requirements) of the 2019 Technical Report (Flagship, ABO).

[40] The process for reviewing items for potential exclusion from score reporting is summarized in section 7.5 of the 2019 Technical Report (Flagship, ABO).

Standard Setting[41]

6a. PLDs [sR]

✓ If the state plans to use New Meridian's policy claims and grade/subject-specific PLDs, the usage is appropriate given the purpose and design of its assessment[42].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ If the state plans modify New Meridian's PLDs or define new PLDs for its assessment, the definition of what it means to be "college and career ready" is similar to that of New Meridian's.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **PLDs**?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar ☐ Does not apply

Reviewer Comments:

[41] The criteria for *Standard Setting* only need to be considered for evaluating readiness comparability.

[42] Policy claims (or policy definitions) are articulated in section 8.2 of the 2019 Technical Report (Flagship, ABO). Grade- and subject-specific PLDs are available for download at the ELA/Literacy and Mathematics pages.

6b. Standard Setting Method [sR]

✓ If the state plans to implement its own standard setting or standards validation process, the type of evidence collected and used to support college and career readiness inferences is similar to that for New Meridian's college and career readiness benchmark (i.e., Level 4)[43].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ If the state plans to empirically map the New Meridian performance levels onto the scale of its assessment, the method for determining the cut scores is technically sound and defensible, especially at New Meridian's college and career readiness benchmark (i.e., Level 4)[44].

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ If the state plans to convene committees to review or validate New Meridian cut scores on its assessment, the design and/or implementation of the standards review or validation process is not likely to lead to different meaning or expectations of the performance levels, especially at New Meridian's college and career readiness benchmark (i.e., Level 4).

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **standard setting method**?

[43] Details of the standard setting process for the New Meridian performance levels are given in section 8.3 of the 2019 Technical Report (Flagship, ABO).
[44] For an example of an empirical benchmarking study conducted to inform the standard setting process, see section 14.5.2 of the 2019 Technical Report (Flagship, ABO).

☐ Highly similar ☐ Adequately similar ☐ Minimally similar ☐ Does not apply

Reviewer Comments:

6c. *Validity of Cut Scores* sR

✓ If the state has conducted or is planning to conduct research studies[45] to validate the cut scores on its assessment, the design and conclusion of the studies are technically sound and appropriate, especially for its readiness (on-track or college and career ready) cut.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

✓ There is evidence[46] supporting the claims, interpretations and uses of the on-track or college and career readiness benchmark on the state assessment.

☐ Agree ☐ Somewhat Agree ☐ Disagree ☐ Insufficient evidence ☐ Does not apply

Reviewer Comments:

Given the preponderance of evidence, how would you rate the degree of similarity between the state's assessments and the New Meridian test forms in terms of **validity of cut scores**?

☐ Highly similar ☐ Adequately similar ☐ Minimally similar ☐ Does not apply

[45] Examples of research studies conducted to validate its cut scores include the benchmarking study, the postsecondary educators' judgment study, and the college and career readiness benchmark validation study.

[46] Evidence of valid interpretation and use of the New Meridian cut scores is provided in sections 14.7 and 14.8 of the 2019 Technical Report (Flagship, ABO).

Reviewer Comments:

Summary

In Table 8 below, please summarize the comparability ratings you gave for each criterion in the previous section and indicate your level of confidence in each rating based on the evidence submitted by the state for the **Scoring** area. If you are not very confident about a rating, please suggest additional evidence that the state could provide to support its comparability with New Meridian test forms for the criterion.

As in the previous section, each criterion includes superscripted symbols that indicates whether evidence of comparability between the state's assessment and the New Meridian test forms is required ("Yes" in Table 7) or recommended ("Recommended' in Table 7). An upper-case "S" or "R" means that evidence of comparability on this criterion is required for scale score or readiness comparability respectively; while a lower-case "s" or "r" means that evidence of comparability on this criterion is required for scale score or readiness comparability, respectively.

Table 8: Summary of Comparability Ratings for the Design Area

Criterion	Comparability rating	Confidence in rating	What additional evidence could be provided? (If somewhat or not confident in your rating)
4a – Scorer qualification [S]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	
4b – Scoring protocols [S]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	
4c – Fairness in scoring [s]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	
4d – Automated scoring [S]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	

Criterion	Comparability rating	Confidence in rating	What additional evidence could be provided? (If somewhat or not confident in your rating)
5a – Psychometric models [S]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	
5b – Scaling and equating [S]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	
5c – Student group analysis [S]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	
6a – PLDs [SR]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	
6b – Standard setting method [SR]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	
6c – Validity of cut scores [SR]	☐ Highly similar ☐ Adequately similar ☐ Minimally similar	☐ Very confident ☐ Somewhat confident ☐ Not confident	

Overall Evaluation

Based on your review of the submitted evidence, what is your response to the *Bottom Line Question* for the **Scoring** area:

*Overall, do you believe that the criteria, methodologies and procedures for scoring the state's assessment with New Meridian content is similar enough (i.e., **adequately similar**) to that of the New Meridian test forms to support **Readiness Comparability**?*

☐ Yes ☐ No ☐ More Information Needed

*Overall, do you believe that the criteria, methodologies and procedures for scoring the state's assessments with New Meridian content is similar enough (i.e., **highly similar**) to that of the New Meridian test forms to support **Scale Score Comparability**?*

☐ Yes ☐ No ☐ More Information Needed

Please provide any additional comments or feedback to support your overall comparability evaluation of the **Scoring** area.
