INTERIM ASSESSMENT QUALITY RUBRIC

Evaluating Quality Claims and Evidence

Advanced Instructional Management: Assessment Literacy
[Email address]

General Directions

The enclosed performance measure rubric is designed to examine the quality characteristics of vendor-made (external) assessments. The rubric is comprised of several technical requirements organized into eight (8) strands. Each requirement is rated on a Likert-type scale ranging from zero (not addressed) to one (fully addressed).

Rater's Task

Step 1. Review information, data, and documents associated with the creation, implementation, and refinement of the selected performance measure.

Step 2. Assign a rating to each component (TASK) within a particular strand using the following scale-

a. (1) = fully addressed
b. (.5) = partially addressed
c. (0) = not addressed
d. (N/A) = not applicable at this time

Step 3. List information, data, and document references supporting each assigned rating.

Step 4. Add notes and/or comments articulating nuances of the performance measure.

Step 5. Compile results for each strand into an overall Summary Matrix.

Summary Matrix

Strand	Points Possible	Points Earned	Strand	Points Possible	Points Earned
Design	5.0		Technical	10.0	
Specifications	5.0		Reporting	5.0	
Development	5.0		Quality	5.0	
Administration	5.0		Costs	5.0	

Total		

Evidence List

Evidence Code	Evidence Name

STRAND 1: DESIGN

Task ID	Descriptor	Rating	Evidence
1.1	The purpose of the performance measure within the system was explicitly stated (e.g., who will be tested, what is the content of the test, what are the uses of the results).		
1.2	A rationale/justification explained the performance measure's design in terms of presentation, format, length, utility, etc.		
1.3	A general description of how the performance measure contributed to the overall performance measure system was provided (e.g., benchmarking student progress between administrations of the statewide performance measure).		
1.4	The performance measure had targeted content standards that represented a full range of knowledge and skills students were expected to master and demonstrate.		
1.5	The performance measure's design was developmentally appropriate for the intended audience and reflects challenging material needed to demonstrate higher-order thinking skills.		
	Strand 1 Summary	***___ out of 5***	
Additional Comments/Notes			

STRAND 2: SPECIFICATIONS

Task ID	Descriptor	Rating	Evidence
2.1	Specification tables articulated the number of items, item types, passage readability, and other information for the performance measure. For a computer adapted performance measure, the specification tables articulated a sampling approach of the targeted content standards.		
2.2	Test blueprints were developed and used to align items to targeted content standards. For computer adapted performance measures, algorithms were used to select items aligned to the targeted content standards.		
2.3	Blueprints/algorithms identified the number and types of items used to measure the targeted content standards and provided information regarding item characteristics (e.g., difficulty, discrimination, cognitive demand, etc.).		
2.4	Items were varied and designed to measure a range of cognitive demands/higher order thinking skills at developmentally appropriate levels in order to reflect the rigor articulated in the targeted content standards.		
2.5	Items were of sufficient quantities and type to sufficiently measure the depth and breadth of the targeted content standards, while providing multiple opportunities for the test-taker to demonstrate content knowledge.		
	Strand 2 Summary	***___ out of 5***	

Additional Comments/Notes

STRAND 3: DEVELOPMENT

Task ID	Descriptor	Rating	Evidence
3.1	Item development committees consisted of subject matter experts for each targeted content area. Committee members developed, modified, and/or reviewed items using standardized guides (i.e., documents the tasks, techniques, and procedures used in creating the performance measure).		
3.2	Item development committees created score keys, including scoring rubrics for human-scored, open-ended prompts (e.g., short constructed response, extended constructed response, writing prompts, performance tasks).		
3.3	Item development committees created and/or reviewed administrative and scoring guidelines.		
3.4	The item development committees examined field tested items in terms of (a) alignment, (b) developmental appropriateness, (c) content accuracy, (d) fairness, (e) sensitivity, and (f) accessibility.		
3.5	Field testing results were used to evaluate the performance of newly developed items. Field testing results were used in making additional corrects/improvements to the item.		
	Strand 3 Summary	***___ out of 5***	

Additional Comments/Notes

STRAND 4: ADMINISTRATION

Task ID	Descriptor	Rating	Evidence
4.1	Response templates/answer documents used an identification mechanism that linked an individual's responses to a particular set of responses. For on-line/PC-based performance measures, test-takers responses can be exported by the test administrator. Administrative procedures outlined steps used to ensure security of personally identifiable information.		
4.2	The performance measure created administration workshops and/or materials to train school test coordinators/personnel. For online/PC-based performance measures, the performance measure developer published procedures used to determine if the capacity exists for an online performance measure (e.g., operating system, required software, etc.).		
4.3	Test administration guidelines included procedures for tracking the distribution and return of testing materials, including guidelines for addressing irregularities during the administration of the test. For online/PC-based performance measures, the performance measure developer provided a detailed description of the steps used to open and close the access portal. This description articulates how test security is maintained during each session.		
4.4	Guidelines contained the step-by-step procedures used to administer the performance measure in a consistent manner. These guidelines addressed procedures such as- • scripts for teachers to orally communicate guidelines • day and time constraints • allowable accommodations • how to distribute and collect performance measure materials • accountability and safeguarding of secure materials For online/PC-based performance measures, ease-of-use by end-users was supported by characteristics such as- • allowing students to split administration sessions		

Task ID	Descriptor	Rating	Evidence
	• having audio capacity for appropriate accommodations • adaptable font sizes • using split screens to keep the passage visible while answering items • saving responses after each item or set of items • splitting the performance measure into more than one session • allowing for non-sequential movement through test items • flagging test items as a reminder to revisit the item • highlight both test items and passages		
4.5	Toll-free telephone and/or web-based support services were available during the administration period of the performance measure. Support services also took the form of guidance documents (FAQs, troubleshooting guides, etc.) that addressed logistical and administrative needs.		
	Strand 4 Summary	***___ out of 5***	
Additional Comments/Notes			

STRAND 5: TECHNICAL

Task ID	Descriptor	Rating	Evidence
5.1	Standard-setting (CRT) or norming (NRT) procedures followed national recognized methods for each subject area and performance measure type. Procedures addressed how performance scores across grade levels allowed for consistent interpretability.		
5.2	For CRTs, performance level descriptors described the achievement continuum using content-based competencies for each assessed content area. Cut scores were established for each performance level. For NRTs, reported scores were based upon the performance measure given their applicable referent group.		
5.3	The performance measure articulates the techniques used to ensure the accurate and reliable scoring of student responses.		
5.4	Human-scored responses had clear and detailed scoring guidelines. Continuous monitoring occurred during scoring, including daily and on-demand reviews of rater accuracy and speed. Scoring quality was maintained by tracking rater scores and reporting inter-rater reliabilities, along with performance on recalibration sets.		
5.5	Analyses were conducted to support item development, fairness/bias evaluations, test construction, item calibration, standard-setting/norming, equating, scaling, and validation activities.		
5.6	Psychometric data showed the results from test scaling and score equating, including the equating of alternative forms. For computer-adapted measures, data showed item exposure rates, blueprint match rates, and other performance statistics.		

Task ID	Descriptor	Rating	Evidence
5.7	Reviews examined the alignment characteristics of the performance measure in terms of: • consistency with specification tables/blueprints • DoK consistent within the targeted content standards • rigor of sampled content • developmental appropriateness • pattern of emphasis		
5.8	The selected equating method was described and shown to be appropriate for the performance measure. Data demonstrated items fit expected parameters with minimal equating error.		
5.9	Reliability coefficients were reported for the performance measure, which included measures of internal consistency. Standard and conditional errors were reported. When applicable, other reliability statistics such as classification accuracy, rater reliabilities, and others were provided.		
5.10	Sources of validity evidence associated with score convergent/divergent with external measures, interrelationship of items and item types, alignment to targeted standards, and equivalency of alternate forms were provided.		
	Strand 5 Summary	***___ out of 10***	

Additional Comments/Notes

STRAND 6: REPORTING

Task ID	Descriptor	Rating	Evidence
6.1	Score reports contained data on how test-takers performed on the performance measure as compared to established performance standards (CRT) or the referent norming group (NRT).		
6.2	Scores were reported using other metrics (e.g., scaled scores, NCEs, PLs) besides raw score points. These scores were explained to non-measurement audiences using jargon-free narratives. For CRTs, the performance level descriptors were available to parents and students within the greater reporting system.		
6.3	Educator reports provided individual student results at the classroom level. These reports included item-level results and showed student performance on different items, along with comparison data (e.g., district, state, normed group, etc.).		
6.4	Interpretative guides were published and made available by the performance measure developer. These guides assisted parents, teachers, and others in understanding the reported performances of students.		
6.5	Parent/student reports provided guidance on score interpretation so parents could address their child's learning needs.		
	Strand 6 Summary	***___ out of 5***	
Additional Comments/Notes			

STRAND 7: QUALITY

Task ID	Descriptor	Rating	Evidence
7.1	Item/operational form review procedures addressed- • sorting/sampling of items based on a balance of targeted content across forms • utilization of item statistics from previously tested items to ensure similar levels of difficulty and complexity • determining the length of the form given the number of pages		
7.2	The performance measure had procedures for editorial control in accordance with professional standards, while ensuring consistency and accuracy of other documents (e.g., administration manuals, directions, response booklets).		
7.3	Testing accommodations reflected those used during regular instruction and/or those afforded during participation on other performance measures. These accommodations did not invalidate the scores.		
7.4	Post administration analyses were used to improve performance measure quality, detect poor item performance, scoring drift, omission rates, and other quality aspects. These results are then used in the upcoming measurement cycle to improve the overall system.		
7.5	Item pools/test banks, norming groups, and operational forms are updated periodically to maintain the relative quality of the performance measure, while minimizing item exposure.		
	Strand 7 Summary	***___ out of 5***	
Additional Comments/Notes			

STRAND 8: COSTS

Task ID	Descriptor	Rating	Evidence
8.1	The total time to administer the performance measure was developmentally appropriate for the test-taker. Generally, this is 30 minutes or less for young students and up to 60 minutes per session for older students (high school).		
8.2	The unit cost (in dollars) was provided for administering the performance measure, including ancillary materials for each test-taker. Scoring costs were provided for the performance measure.		
8.3	Scored measures were available to educators and students/parents in a timely manner. For self-scored or computer adapted tests, results were available within 48 hours. For externally scored measures, results were made available within five (5) workdays.		
8.4	The overall performance measure costs included multiple forms for pre/post administration, interpretative guides, scoring, and reporting services. Costs for secondary vendors (if required) were made readily available for comparison purposes.		
8.5	The performance measure did not require additional fiscal resources (e.g., licensing fees, set-up fees, etc.) prior to implementation. All scored results became the property of the educator/district.		
	Strand 8 Summary	***___ out of 5***	
Additional Comments/Notes			