1. Fairness

Please demonstrate how you assure fairness in your assessments.

1. Statistical evidence of fairness in development
2. Fairness in development (Universal Design)
3. Fairness in administration; accommodations (accessibility)
4. Fairness in administration; standardized directions
5. Fairness in administration; practice items

Rubric item description	Above Good	Good	So-So	Not OK	Comments
	3	2	1	0	
Statistical evidence of fairness in development	Differential Item Functioning Analysis (DIF) used to refine items on test, with appropriate sample	Test development included all subgroups (with appropriate sample and stratified random sample)	Test development population was diverse but not specified further	Little or no evidence of empirical analysis of fairness/bias provided	How did the developer attend to bias using statistical methodology? Did they at least include all subgroups while testing the function of items? Did they directly examine item function for each subgroup (DIF analysis)?
Fairness in development (Universal Design)	Process described for reviewing item wording, visuals for accessibility, cultural bias, offensive content, visual distractors in items and directions (all present) Exemplars provided	Process described for reviewing item wording, visuals for accessibility, cultural bias, offensive content, visual distractors in items and directions (majority present) Exemplars provided	Items and directions reviewed by experts for bias/fairness, specific details not provided	No evidence of expert review of fairness/bias provided	What does the developer tell us about processes used to examine and minimize bias? There are a variety of standard methods used while developing and refining items, tests and other materials. These are often known as principles of universal design. These help insure accessibility of test items.
Fairness in administration; accommodations (accessibility)	Validated, feasible accommodations are described, including for example, sensory impairments,	Feasible and appropriate accommodations are described, including for example, sensory	List of accommodations	No evidence of guidance for accommodations	Does the test include specific information about acceptable accommodations for students who may have disabilities such as visual or hearing impairments where feasible and appropriate? Validation means that there have been studies of the effects of

	limited English proficiency	impairments, limited English proficiency			these accommodations, with appropriate cautions about risks to data use.
Fairness in administration; standardized directions		Administration is standardized		No evidence of standardized administration	This is a yes/no item. Either administration directions are standardized, or they are not. Standardized instructions help insure fairness in administration.
Fairness in administration; practice items	Practice items include sample questions that represent all item structures that will be present on the assessment at each grade. Samples of relevant accommodations presented when appropriate	Availability of practice items at each specific grade. Samples of relevant accommodations presented when appropriate	Practice test items for grade spans	No practice test content available	This is about the question structures - do the items create a barrier to measuring knowledge and skill? Practice items provide students with an opportunity to experience the item structures that will be present on the test. Accommodations are defined in practice items where appropriate.

2. Availability

Please demonstrate the availability of your assessments.

6. Grade availability
7. Availability in last quarter of school year

Rubric item description	Above Good	Good	So-So	Not OK	Comments
	3	2	1	0	
Grade availability	3-11 + additional grades	grades 3-11		Not all of 3-11	Legislation requires grades 3-11
Availability in last quarter of school year		Yes		No	Legislation requires spring administration

3. Describes Achievement

Please demonstrate how your assessments accurately describe student achievement.

8. Accurately describes student achievement
9. Accurately describes growth

Rubric item description	Above Good	Good	So-So	Not OK	Comments
	3	2	1	0	
Accurately describes student achievement	Materials indicate reporting at claim and standard levels as well as DOK. Reports available at the system, building, section, classroom and student level.	Materials indicate reporting at claim and standard levels as well as DOK. Reports available at the system, building, and student level.	Materials indicate reporting at claim and standard levels as well as DOK. Reports available at the system and building level.	Materials indicate reporting at claim and standard level but not by DOK.	To what level of detail can the reports that the test provides answer questions at fine and coarse grain level? Is Depth of Knowledge (DOK) represented (AKA: rigor and cognitive complexity)? Do the reports include individual and group information?
Accurately describes growth	In addition to what is in #2, scaling includes flexibility to measure growth and achievement above and below the targeted grade level or ability to accurately/precisely place a student performance on a continuous, cross-grade achievement scale. Forms are equated	Interpretable across grades/years, has data on growth using a scaled score that is vertically articulated across grade levels by each domain, at the systems, building and student level. Forms are equated	Interpretable across grades/years, has data on growth using a scaled score that is vertically articulated across grade levels by content area (R & M), at the building and student level. Forms are equated	No vertical scaling	This is about the test construction and scaling. Did they develop a calibrated vertical scale that will allow us to compute change scores across years that can be interpreted as growth? Does the scaling of the test allow flexible testing on lower or higher content for students at the upper and lower limits of achievement, where floor and ceiling effects of many tests limit interpretation? Were efforts made to equate forms across years and/or administrations to provide stable interpretation (OR, if adaptive test, to ensure equivalence of tests)?

4. Validity

Please demonstrate how you have determined your assessment is valid

10. Criterion validity coefficient (correlational evidence)
11. Quality of Validity evidence

Rubric item description	Above Good	Good	So-So	Not OK	Comments
	3	2	1	0	
Criterion validity coefficient (correlational evidence)	>0.8	>0.70	>0.60	<0.6 or none reported	Does this test measure the same thing as other tests?
Quality of Validity evidence	Multiple studies using criteria from level 2 with different populations and including all applicable statistics	Validity evidence using correct methodology with reasonable samples (size and representativeness). Comparison measures used are of reasonable technical quality and measure the desired content/constructs	Validity evidence using correct methodology, but marginal samples (size and/or representativeness). Comparison measures used measure the desired construct, but are not "mainstream" assessments and the samples used for the comparison measure's technical data are not representative (a sample of convenience).	Correct methodology, but poor or dated sample	Correlation measures that estimate validity should use assessments that are themselves reliable and valid, and measuring the desired construct - in other words, we are comparing reading tests with reading tests. Less desirable data may be found where the assessment used to validate is a less-than robust measure or use samples that are far from representative in either population, or timeframe. Further evidence of validity may be found in the alignment items.

5. Reliability

Please demonstrate how you have determined that your assessment is reliable

12. Internal consistency (alpha, split half, marginal)
13. Stability over time (test retest, alternate form)
14. Scorer consistency (inter-rater agreement in some form) (if applicable)
15. Quality of reliability evidence

Rubric item description	Above Good	Good	So-So	Not OK	Comments
	3	2	1	0	
Internal consistency (alpha, split half, marginal)	>0.9	>0.8	>0.7	<0.7	
Stability over time (test retest, alternate form)	>0.9	>0.8	>0.7	<0.7	
Scorer consistency (inter-rater agreement in some form)	>0.9	>0.8	>0.7	<0.7	If applicable (i.e., scoring of the test requires human or machine scoring of student constructed response, where scorer error could be a factor). If not applicable, this item will not count for or against in the score.
Quality of reliability evidence	Multiple studies using different populations and including all applicable reliability statistics. Includes correct methodology as in #2.	Reliability evidence using correct methodology with reasonable samples (size, relevance to Iowa students and, representativeness) .	Reliability evidence based on limited or non-representative populations	Reliability evidence on previous or related version	We are looking for sufficiently thorough reviews of reliability, with all applicable statistics. Not only are the reliability values important, but it is important for us to make some judgment about the methods used to gather the results, and how those values were calculated. High values resulting from low quality research data are suspect. We want to be certain the values reported are meaningful and interpretable in the Iowa context.

6. Piloted/Tested in Iowa

Please demonstrate that your assessment has been piloted in Iowa

16. Piloted in Iowa (item tryout)
17. Tested in Iowa (field tested)

Rubric item description	Above Good	Good	So-So	Not OK	Comments
	3	2	1	0	
Piloted in Iowa (item tryout)		yes		no	Piloted means item tryouts and small sample test groups Pilot Test: A stand-alone administration of test items, tools or a system, to evaluate how particular items function prior to a field test and operational use. The pilot test generally occurs with a sample of students that matches the purpose of the pilot.
Tested in Iowa (field tested)		yes		no	Field Test means larger scale testing of a fully developed test Field Test: An administration of the field test to evaluate how the test functions prior to operational use. This generally occurs after a pilot test, using a significantly larger, more representative sample of students than a pilot test

7. Alignment

Please demonstrate that your assessment is aligned in the following ways

18. Methodology of content alignment to domains, standards and clusters
19. Tables of specifications
20. Amount of content coverage
21. Evidence of alignment in Depth of Knowledge (DOK) (AKA rigor or cognitive level)
22. Language is consistent with the Iowa Core

Rubric item description	Above Good	Good	So-So	Not OK	Comments
	3	2	1	0	
Methodology of content alignment to domains, standards and clusters	Post hoc alignment study based on content in #2, PLUS evidence from #2	Methodology of test construction clear about content alignment specific to domains, standards and clusters and describes methodology for confirming alignment of items during test development	The domains match, but not the standards or clusters (next level down)	Reorganization or relabeling of an existing assessment to match Iowa Core content. —OR— No evidence provided	How much care was placed in test development relative to alignment with the Iowa Core? This item represents the internal processes used during test construction and the level of detail addressed by those processes. Note that the Iowa Core uses different terms to categorize ELA and mathematics. For efficiency in this document, we use “domains” to mean the largest components (reading, writing, etc.) and use “standards” and “clusters” to mean progressively finer-grained categories within the domain.
Tables of specifications		Tables of specifications for the developed or new test are provided.		No table of specifications provided	The tables of specifications contain the blueprint for test construction, including content, DOK, relative emphasis, etc.

Amount of content coverage	N/A	All domains presented, majority of standards and clusters presented	All domains presented but fewer than half of standards are identified	Most domains presented — or — no evidence presented	How much of the Iowa Core does the test purport to measure?
Evidence of alignment in Depth of Knowledge (DOK) (AKA rigor or cognitive level)	N/A	Summary table detailing the distribution of DOK levels on items	Describes how DOK is addressed in test construction methodology. DOK described in general terms	No mention of DOK	We are looking for evidence that depth of knowledge was addressed systematically during test development.
Language is consistent with the Iowa Core	N/A	Displays, reports, and technical information match the language of Iowa Core domains and standards and DOK as defined in the Iowa Core	N/A	Inconsistent, incomplete, or no match with terminology of Iowa Core	This will be a word match of samples - do the words that are used in the test documents match the words used in the Iowa Core? Can people use this without translation cards?

8. College/Career

23. Please demonstrate that your assessment measures progress toward college or career (content) readiness

Rubric item description	Above Good	Good	So-So	Not OK	Comments
	3	2	1	0	
Ability to measure progress toward college or career (content) readiness		Statistical analysis measures progress toward college or career (content) readiness indicators.		No college or career statistical analysis available	Does the test offer a means to measure progress toward college or career (content) readiness? College/Career Content readiness relates directly to mastery of Iowa Core content (see also alignment rubric).

9. Technical Supports

Please demonstrate the technical supports that are available

24. Training on assessments and interpretation of reports
25. Availability of results - machine scored (including AI scored constructed response items)
26. Availability of results - human scored (student constructed responses)

Rubric item description	Above Good	Good	So-So	Not OK	Comments
	3	2	1	0	
Training on assessments and interpretation of reports	Online modules for key elements, with instructional materials on administration and report interpretation (as in #2 and #1)	Instructional materials on administration and report interpretation (and manuals as described in #1)	Technical and user manuals with administration and interpretation information	Incomplete materials for administration and interpretation, no technical manual	Multiple training methods are desired, with flexible supports for key elements. Self-paced training options desired.
Availability of results - machine scored (including AI scored constructed response items)	Individual results available in real-time. Classroom, building and system available within 24 hours of last testing (receipt of student responses)	Group and individual reports available within one week of test completion (receipt of student responses)	Group and individual reports available within 2-3 weeks of receipt of student responses	Group and individual results available in more than 3 weeks of receipt of student responses	Level 3 assumes electronic assessment and reporting capability. Other levels allow for either electronic or paper administration and reporting.
Availability of results - human scored (student constructed responses)	Group and individual reports available within three weeks of receipt of student responses	Group and individual reports available within one month of receipt of student responses	Group and individual reports available within six weeks of receipt of student responses	Group and individual reports not available within six weeks of receipt of student responses	Used if human scoring required, split points between machine and human scoring.