Using Validity Arguments to Evaluate the Technical Quality of Local Assessment Systems

Chad M. Gotch
Washington State University

Marianne Perie
National Center for the Improvement of Educational Assessment

Paper presented at the annual meeting of the American Educational Research Association conference. April, 2012, Vancouver, BC, Canada.

Abstract

Despite the centrality of validity concerns to the practice of educational assessment, little practical guidance exists for creating and evaluating validity arguments. This paper proposes a validity argument framework for evaluating the technical quality of locally-developed high school end-of-course examinations that are intended to supplement or supplant state-level examinations. The framework seeks to balance concerns for psychometric rigor with feasibility of implementation. Essential elements of contemporary validity theory have been adopted and translated into accessible terminology and a practical validation process. The State of Pennsylvania provides an exemplar assessment context for the implementation of the validity argument framework. Examples of required validity evidence and claims to support the locally-developed assessment are provided.

Using Validity Arguments to Evaluate the Technical Quality of Local Assessment Systems

Validity is the cornerstone upon which public confidence in high-stakes testing programs is built. A sound validity evaluation is a necessary component of any kind of high-stakes assessment use in the schools; however, such a framework may often be lacking in assessments developed at the local level (e.g., school district; Sperling & Kulikowich, 2009). This paper will illustrate a practical application of forming and evaluating a validity argument as a basis for evaluating locally-developed assessments, and highlight the tensions among current validity theory and practical implementation issues, especially at the local level.

Several states have encouraged or even mandated that local school districts create their own assessment systems, and use them in place of the state-developed test. These local assessment systems have served a range of purposes, but often have been tied to student graduation determinations, clearly a high-stakes use. Given such purposes, it reasons that state assessment leaders and policy makers are concerned local systems have appropriate degrees of technical quality. Typically, locally-developed tests need to undergo a technical quality evaluation to be certain that they are appropriately aligned with the state-mandated content and that the interpretation of the performance levels remains the same as the state-level test (if applicable). Evaluating technical quality, however, can be a herculean task for a school district that lacks the benefit of a fleet of personnel with advanced educational measurement training. We argue that evaluations of technical quality should be couched within a validity evaluation framework. While several quality academic writings on validity theory exist (e.g., American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Kane, 2006; Lissitz, 2009), few resources exposit detailed guidance for creating and evaluating a validity argument. Furthermore, a similar dearth of resources leaves little guidance to state departments of education charged with overseeing a validity evaluation process across a confederacy of school districts. In this paper we discuss tangible issues regarding the implementation of a local assessment evaluation system, and provide real-world insight for wrestling with these issues.

The purpose of this paper is to explore the demonstration of technical quality in a local assessment system by using a validity argument framework. Previous efforts have introduced validity argumentation to alternative assessment systems for students with disabilities (Marion & Pellegrino, 2006; Marion & Perie, 2009). This presentation goes a step further by addressing the use of validity arguments in local assessment systems that are intended to supplement or supplant state systems.

Validity in the context of local assessment systems

Validity is defined as the “degree to which evidence and theory support the interpretations of the test scores entailed by proposed uses of the test” (AERA, NCME, & APA, 1999, p.9). Messick (1989) discussed gathering validity evidence to evaluate intended test score interpretations, and categorized these forms into five categories—test content, associations with other variables, test structure, response processes, and consequences of testing. The forms of evidence advanced by Messick, have become widely adopted across the educational measurement field and are reflected in *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). Classification of these forms of evidence and the description of how they contribute to an overarching concern for accurate representation of a construct marked an important step in the evolution of validity theory, but little guidance was provided on how to assemble disassociated acts of gathering evidence into a coherent whole.

The argument-based framework for validity (e.g., Kane, 2002, 2006) addressed this need by framing validity in terms of a chain of logical statements. Toulmin, Rieke, and Janik's (1979) model of reasoning conceptualizes the formation of an argument as the piecing together of a sequence of claims and reasons that support one's position. The six elements of a single logical link in the chain of arguments are claim, grounds, warrants, backing, modal qualifiers, and possible rebuttals. The logic process moves from grounds (i.e., data or observations) to claims, with the other elements providing context and delimitations. Single arguments can be assembled into a larger overarching argument, as the claim from one argument becomes the grounds for another argument.

Collectively, this chain provides contextualized explanations of test scores. Applied to the evaluation of a local assessment system, we must consider the inferences and explanations we intend to make. What are the intended claims? Such considerations are shaped by the specific legal code driving the local assessment context. At the heart of all considerations, however, is the desire to make some claim about the quality and accuracy of scores (or other forms of results) coming from the assessments. Interpretive arguments in the model of reasoning outlined above allow for judgment of the extent to which the local system maintains appropriate levels of quality and accuracy.

The case of Pennsylvania

Pennsylvania is a state with a tradition of strong control residing in local school districts. This tradition was maintained when state code was amended in 2010 to establish a new set of high school graduation requirements. While the state is developing new end-of-course assessments to test students' competencies in core curricular subjects, the Keystone exams, a provision was written into law that allowed students to demonstrate competency on locally-developed and independently-validated assessments. This provision was an extension of previous policy that allowed local assessment systems for non-state-mandated content areas. Now, local municipalities will be able to implement their own assessments in content areas considered core for graduation purposes.

The challenge of this new provision will be to ensure that inferences of student performance on local assessments are comparable to inferences drawn from the Keystone exams. Investigation of the current model of local assessments revealed substantial issues surrounding standards of proficiency. In order to render fair decisions on students and instill public confidence in the state's education system, it is imperative to establish an evaluation process for the local assessments. To achieve such an end, the state established a Local Assessment Validation Advisory Committee (LAVAC) comprised of representatives from the Department of Education, State Board of Education, School Boards Association, and local district leaders. The charge of the committee is to develop the criteria for the local assessment validation process and for the selection of approved evaluators. This context provided an opportune chance to explore the development and evaluation of validity arguments in a practical setting.

Development of an argument-based validity evaluation

Artifacts from previous ventures into local assessment systems by other states, the meetings of a state local assessment validation advisory committee, and cases of developing validity arguments for other assessments (e.g., TOEFL, Chapelle, Enright, & Jamieson, 2010) served as primary sources of reference. Establishing criteria for technical quality in a local assessment system required constant consideration of feasibility while adhering to the mandates of a sound validity argument for judging student proficiencies. Given limited resources common to local school districts, framing a validity argument resource for local assessment systems focused on front-end design considerations, placed in a comprehensible context that demonstrates how sound design leads to defensible conclusions.

To move from abstract pontificating about score inferences to concrete guidance in assembling a validity argument, we advance a framework for specifying interpretive arguments that reveal aspects of technical quality. The framework (Figure 1) is based on a simplification of the full model of reasoning advanced by Toulmin et al., and adopted by Kane for the educational measurement field. The full model was simplified to maintain feasibility of implementation in settings lacking human and intellectual capital, while retaining the core elements of the argument framework. Claims reflecting necessary precursors to valid score interpretations (e.g., *The local assessment system maintains an adequate level of rigor*) provide the structure. Each of the claims is further explicated by the statements in the evaluation criteria matrix (see Appendix A). The evaluation criteria are meant to help the districts determine what type of evidence is needed. Figure 1 derives from the general principle underlying Kane's (2006) work, which asks us to provide the data, warrant, and claim. That is, we start with data and determine how it provides evidence to support a claim. Local school districts, with the assistance of exemplars of supporting evidence (see Appendix B), are tasked with providing the evidence and backing for these claims. This framework allows system evaluators to render judgments based on the clarity, coherence, and plausibility of the interpretive arguments.

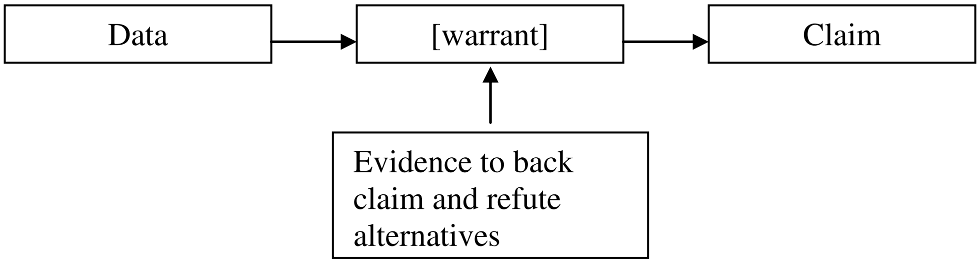

Figure 1. Argument framework for local assessment validation

To assemble an overarching, cohesive argument about the quality of a local assessment, we worked backwards from the proposed interpretations to develop a series of claims and assumptions that must be true for the interpretation to be valid. We then provide evidence and data to support each assumption and claim to show that our proposed interpretations are valid. This approach asks test developers to think of reasons why the intended inferences might not be supported. All this means that we basically state what we think the test does and then try to disprove it. Consider, for example, the statement “an increase of student scores reflects a greater understanding of the content.” An alternative hypothesis could be that “an increase in student scores reflects greater teaching of test-taking strategies.” Essentially these contrasting explanations ask the question, if student scores are higher in the second year than in the first, does that mean that they know more or that they and their teachers are more familiar with the tests? Collecting evidence both to refute the second claim as well as to support the first would strengthen the validity evaluation. We want to prove that students actually know more but we prove it by trying to disprove it or the alternatives. If we disprove the alternatives and cannot disprove our desired assumption, then we have evidence that the test does what we say it does. In practice, it is not possible to search for all the reasons, but the framework provides us guidance for developing studies that refute other possible explanations for a finding.

Shown below (Table 1) are the proposed interpretations and claims for the Pennsylvania Local Assessment System. Two top-level interpretations were identified—one related to the quality of data provided by the local assessment, and one related to the relative rigor of the local assessment. The common sources of validity evidence were re-framed into a language consistent with what was familiar and of concern to test users. The categories used in this framework were alignment, fairness, establishment of proficiency levels, and consistency. Each category contained two claims that would lay the foundation for the proposed interpretations. Figure 2 demonstrates the flow of claims to proposed interpretations.

Table 1. *Proposed interpretations and claims for the Pennsylvania Local Assessment System*

Proposed interpretations	1. The local assessment provides data on students' readiness for college or careers that is equally good or better than the Keystone Exams. 2. Proficiency scores on the local assessments are equally as or more rigorous than proficiency scores on the Keystone exams and cover equivalent material.
Alignment claims	• The items on the local assessment represent the content standards to the same breadth and depth as the Keystone items. • The content coverage of the local assessment is aligned with the Keystone assessment.
Fairness claims	• Test scores across all identifiable and relevant student groups have comparable interpretations with respect to the course content area. • All identifiable and relevant student groups receive equitable treatment within the assessment system.
Establishing proficiency levels claims	• The local assessment system maintains an adequate level of rigor in the proficiency levels. • Judgments of student proficiency are set using a researched and established methodology.
Consistency claims	• Student scores do not depend upon assignment to a particular scorer, test form, school, test-taking location, or test-taking year. • Student scores are reliable indicators of achievement in the course content area.

The goal of each validity evaluation submission is to provide evidence for each of these claims. Thus for each claim, the submitter should provide evidence in the form of testing documents (e.g., sample tasks, test blueprints, instructions, policies) or study results (internal or external) and an explanation of how that evidence supports the claim. Again, to strengthen the evaluation, evidence refuting alternative hypotheses will strengthen the application.

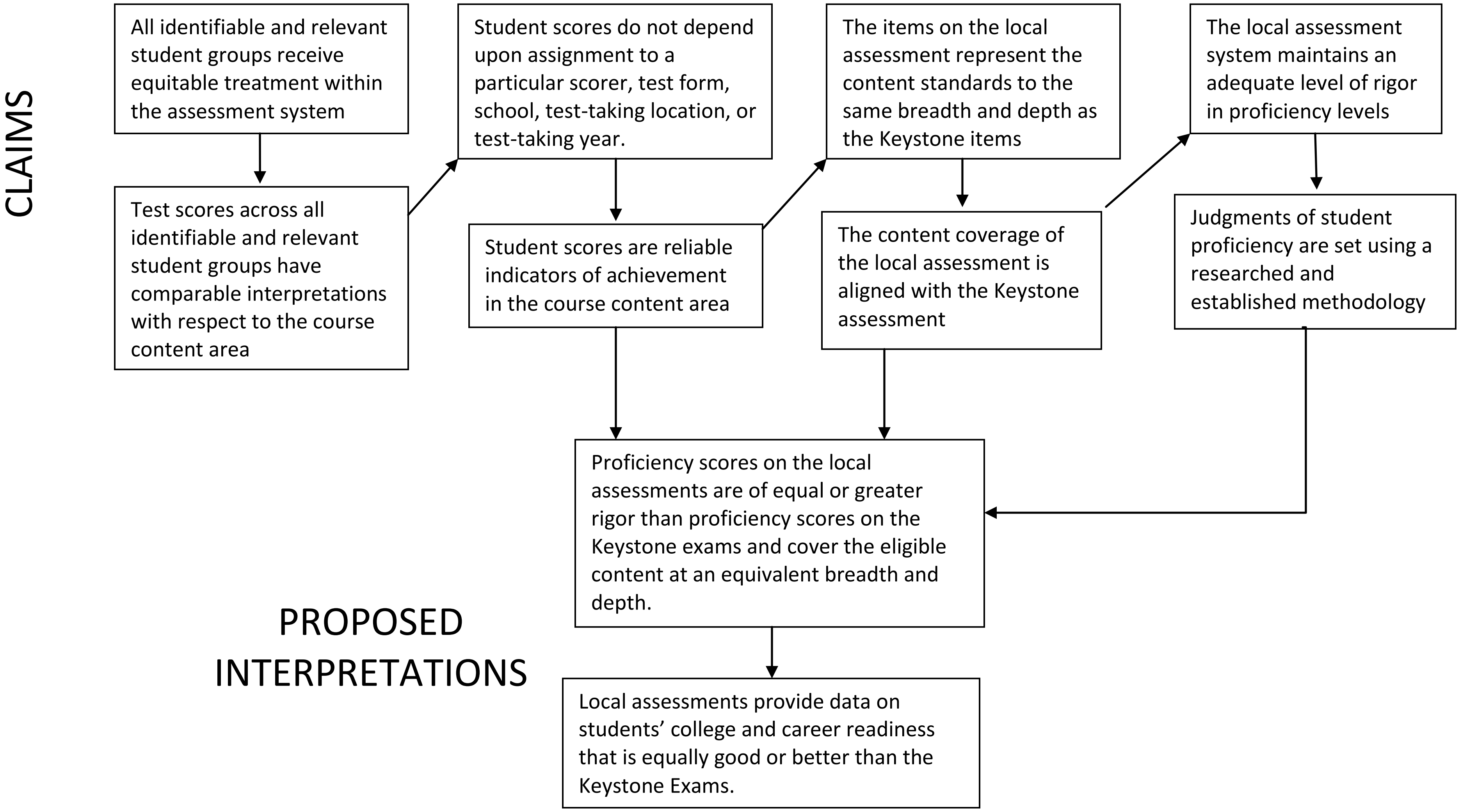
CLAIMS
All identifiable and relevant student groups receive equitable treatment within the assessment system
Test scores across all identifiable and relevant student groups have comparable interpretations with respect to the course content area
Student scores do not depend upon assignment to a particular scorer, test form, school, test-taking location, or test-taking year.
Student scores are reliable indicators of achievement in the course content area
The items on the local assessment represent the content standards to the same breadth and depth as the Keystone items
The content coverage of the local assessment is aligned with the Keystone assessment
The local assessment system maintains an adequate level of rigor in proficiency levels
Judgments of student proficiency are set using a researched and established methodology
Proficiency scores on the local assessments are of equal or greater rigor than proficiency scores on the Keystone exams and cover the eligible content at an equivalent breadth and depth.
PROPOSED INTERPRETATIONS
Local assessments provide data on students' college and career readiness that is equally good or better than the Keystone Exams.

Figure 2. The flow from claims to proposed interpretations about local assessments in the Pennsylvania Local Assessment System

To submit a validation of intended inferences to be drawn from a locally-developed assessment, municipalities may organize their evidence using a template as the one provided in Table 2. The table contains three columns—one each for the data, explanation of how it supports the claim (or refutes an alternative hypothesis), and the claim the evidence supports—with only the third column completed. The district will be responsible for completing the first two columns. We demonstrate this process for the alignment claim, *The items on the local assessment represent the content standards to the same breadth and depth as the Keystone items*. Evidence could include training materials for item writers and item reviewers as well as an external alignment study. Ideal evidence would demonstrate that that the items are fully aligned with the assessment anchors and that all of the assessment anchors are covered by the local assessment. Provision of this evidence would also counter the claim that certain anchors were omitted or reduced in importance compared to what was intended with the Keystones. An example of a completed template for alignment claims can be found in Appendix C. (In the full handbook, available at http://websites.pdesas.org/localassessmentvalidation/2011/10/13/337018/page.aspx, templates also exist for claims around fairness, establishing proficiency levels, and consistency.) In addition, the actual instructions for item writers and reviewers would be included as well as the full report from the external alignment study.

Table 2. *An example template for providing evidence and backing for an alignment claim*

Data or Evidence	Explanation of how it supports the Claim	Claim
		The items on the local assessment represent the content standards to the same breadth and depth as the Keystone items.

Conclusion

The validity argument framework has not been attempted widely in statewide local assessment systems. A benefit to the validity argument framework is that it allows for evaluations to be made in the absence of a strong, unified theory of an underlying construct (Chapelle, Enright, & Jamieson, 2010). Such a situation describes the context of much educational testing. In the case of local school districts implementing an assessment system, the validity argument framework provides clear guidance, and leads local professionals toward thinking more deeply about the purposes of the assessment system and the assumptions upon which judgments about students depend. Adoption of an argument-based validity framework for the evaluation of local assessments may facilitate the cooperation of local and state policymakers to implement a process that can both meet legislative requirements and clarify the burden of proof required of local practitioners.

From a theoretical perspective of validity, the argument-based framework proposed for locally-developed assessments begins to answer the fundamental question of the "degree to which evidence and theory support the interpretations of the test scores entailed by proposed uses of the test". Thus, the framework embodies and puts into action the most contemporary notions of validity theory. By extending this framework to a real, high-stakes setting the framework fills a scholarly void of turning best theory into best practice. The proposed framework could enlighten educational professionals to best practices in validation, resulting in assessment systems that instill public confidence.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*, 3-13.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, *21*, 31-41.

Kane, M. T. (2006). Validation. In *Educational Measurement*, American Council on Education / Praeger Series on Higher Education (4th ed., pp. 17-64). Westport, CT: Praeger Publishers.

Lissitz, R. W. (2009). *The concept of validity: Revisions, new directions, and applications*. Charlotte, NC: Information Age Publishing.

Marion, S. F., & Pellegrino, J. W. (2006). A validity framework for evaluating the technical quality of alternative assessments. *Educational Measurement: Issues and Practice, 25*: 47-57.

Marion, S., & Perie, M. (2009). A validity framework for evaluating the technical quality of alternative assessments. In W. D. Schaeffer & R. W. Lissitz (Eds.), *Alternative assessment: Proceedings from the 8th annual MARCES conference* (pp. 113-125). Baltimore, MD: Brookes Publishing.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Sperling, R. A., & Kulikowich, J. M. (2009) *Local Assessment Validity Study Report*. Retrieved from http://www.ed.psu.edu/educ/lavs/lavs-full-report

Toulmin, S. E., Rieke, R. D., & Janik, A. S. (1979). *An introduction to reasoning*. New York: Macmillan.

Appendix A: An criteria matrix for evaluating alignment claims for the Pennsylvania Local Assessment System

Evaluation Criteria	Superior	Satisfactory	Insufficient
Alignment	In addition to the evidence characterizing the satisfactory level: • Evidence of depth of knowledge alignment from results of "think-aloud" protocols or other similar analyses • Evidence from an external alignment study • No gaps in coverage of the standards, all items/tasks are aligned to specific standards, and depth of knowledge represented by the items/tasks matches the expectations for depth of knowledge in the standards	• Documentation of adequate sampling of all content standards • Evidence from an internal alignment study that used a two-way alignment process • Few gaps in the coverage of the standards, all of the items/tasks are aligned to specific standards, and there is a range of depth of knowledge (including DOK 4) represented by the items/tasks • Plans for periodic review of alignment	• Items represent content standards, but many standards are unaddressed • The content standards are represented well, but the depth of knowledge required to correctly answer items is not in alignment with the standards

Appendix B: Exemplar evidence and backing for an alignment claim in the Pennsylvania Local Assessment System

Alignment

Data or Evidence	Explanation of how it supports the Claim	Claim
Include specific evidence that shows that the items are aligned to the course content standards. This evidence could include: • ***Test blueprint or specifications*** • *Item specifications* • ***Written instructions for item writers*** • ***Written instructions for item reviewers*** • *Sample tasks of high DOK* • ***Alignment study done by district or school staff*** o ***Technical report explaining process and results*** o ***Matrix of items to course content standards*** • *External alignment study* o *Technical report explaining process and results* • *Research studies examining the constructs tested by each item, such as a cognitive lab or think-aloud studies*	*Describe in words how the evidence submitted shows that the local assessment matches the course content standards (assessment anchors and eligible content) and/or the test blueprints for the Keystones. Be sure to discuss matching both the breadth and depth of knowledge of the target course content standards. You need to show that you've matched EVERY content standard with sufficient balance of representation and are testing a range of depth through Depth of Knowledge Level 4.* *If using a unique testing approach, showing the content measured might require additional evidence such as an external alignment study or research evidence of student knowledge tapped by the assessment.*	*The items on the local assessment represent the content standards to the same breadth and depth as the Keystone items.*

Appendix C: A simple example of a submission to address an alignment claim in the Pennsylvania Local Assessment System

Note: This submission would judged as satisfactory, assuming the attachments provided information that fully met the evaluation criteria

<table>
<tr><th colspan="3">Introduction:</th></tr>
<tr><td colspan="3">We created a local assessment for our students in Chemistry for several reasons. One, we wanted to maintain control over the integration of the assessment score into the grade. Second, we feel that a strong science assessment should include a performance component where students integrate their understanding of scientific procedures with their knowledge of a specific content area. Therefore, our assessment contains 25 multiple-choice items focused on the content knowledge, 5 open-ended items that focus more on the process, and 1 performance task that requires students to implement scientific procedure to investigate an issue in the chemistry content area. Students who choose this subject as one of their graduation requirements must score Proficient or above to graduate.</td></tr>
<tr><th colspan="3">Alignment</th></tr>
<tr><th>Data or Evidence</th><th>Explanation of how it supports the Claim</th><th>Claim</th></tr>
<tr><td>Test Blueprint (Exhibit A1)</td><td>Blueprint shows the distribution of items across assessment anchors with target depth noted.</td><td rowspan="4">The items on the local assessment represent the content standards to the same breadth and depth as the Keystone items.</td></tr>
<tr><td>Instructions for item writers (Exhibit A3)</td><td>The instructions clearly show that item writers were to develop items such that our item pool contains items of a similar or greater breadth and depth as the Keystones, while maintaining a similar balance of representation.</td></tr>
<tr><td>Instructions for item reviewers (Exhibit A4)</td><td>The item reviewers were asked to independently rate the assessment anchor and depth of knowledge assessed by each item. These ratings were then compared to the original targets to be sure the final item pool contained items of similar or greater depth and breadth and at a similar ratio as the Keystones.</td></tr>
<tr><td>External alignment study – see the technical report with final results (Exhibit A5)</td><td>The two-way alignment study completed by an independent contractor shows that the local assessment is fully aligned with the eligible content and the balance of representation is similar to that of the Keystones.</td></tr>
</table>