

## **Developing Reliable and Valid Assessments**

### **Overview**

Most faculty members have been teachers for many years and have been involved in developing assessments related to Alabama State Standards. Hence the basic process of developing an assessment should be familiar to faculty members. This help module is primarily designed to review the process and perhaps help you to reflect on the assessments that you have developed to meet the Alabama Quality Teaching Standards (AQTs), Alabama State content area standards in teaching and other fields, and standards in program areas that are not part of the certification process. Below, is a description of aspects of the process that you should take into account or should have taken into account when developing assessments.

### **What is the Nature of Standards (and Indicators)?**

Probably the first challenge for developing assessments is to understand the unit of assessment that you will be working with in designing your assessment. The term standards can refer to a variety of statements about competencies that students have in a domain. Alabama Quality Teaching Standards are organized in a three level format. There are the standards, indicators of competence on those standards, and knowledge and ability statements that reflect competence on the indicators. For the Alabama Quality Teaching Standards, and other standards, it is the knowledge and ability statements that will be the unit of assessment. Scores are required on each of the knowledge and ability statements that are part of the standards. When standards or indicators are referred to here, we will talk about them as synonymous with the knowledge and ability statements when discussing state teaching standards. Standards outside of the teaching and support fields have different organizational structures depending upon the field. Another important distinction to discuss is that standards are not necessarily synonymous with instructional objectives. Several objectives (sometimes even a set of unrelated ones) may appear as part of a standard, indicator, knowledge or ability statement. Hence when choosing what to assess, you may find you will have to assess more than one artifact for an indicator. On the other hand, you may also find that you can use one artifact to assess several indicators or parts of an indicator.

As noted above, Alabama State Standards are typically either stated as knowledge standards or ability statements. Knowledge statements generally reflect what we typically learn in academic courses and can sometimes be assessed in standard ways (like tests and papers). However, some knowledge statements involve application, so performance assessments may be in order to assess them. Ability statements, on the other hand, are likely to require performance-based evidence that a student can carry out a task. Hence, they are often assessed using artifacts or observations from field experiences. There are some ability indicators where a well designed simulation might be a viable alternative as well.

## **How do I translate the Indicator/Standard into an assessment?**

There are several approaches to designing instruction and assessments that are effective. One strategy is to use already existing assessments and then develop a scoring scheme for translating the assignment grading into the four point state rubric. If you do not have an existing assessment, then you will need to design one. Two approaches that work very well are Merrill's (2002) [First Principles model](#) and Wiggins and McTighe's (2005) [understanding by Design model](#). Both of these models focus first on what you expect the learner to do once he or she is out in the real world. The models focus on the essential understandings necessary to show competence, and emphasize building learning tasks and assessments that are authentic indicators of the knowledge in the context that the knowledge will be applied.

Since many of the ability statements will be assessed in field experiences, the task here may be simply making sure that there is opportunity to observe the particular indicator in the context of that field experience. The questions that need to be asked concern the adequacy of the observation. Is it something that can be successfully measured on one occasion? Does it need multiple observations? Do you need to make sure that there is a systematically developed context for evaluating that particular skill? Are there products or artifacts that are generated that can be used to either evaluate success or supplement observation? The more systematic the observation, the easier it will be to score the observation. However, caution must be taken to make sure that changing the observation context too much does not create a contrived or less ecologically valid situation.

As noted above, simulations can sometimes be developed that can be used to assess standards/indicators. Simulations can sometimes make the assessment easier to do within the context of a class, but care must be taken to make sure that the simulation retains its ecological validity. The major concern with simulations is that they have to be done in such a way that their outcomes would be unlikely to look different if the student were in the everyday world of teaching and learning. A recent text by Thornton and Mueller-Hanson (2004) presents a thoughtful discussion about the validity concerns associated with developing simulations. Although it examines simulations from the point of view of industrial-organizational psychology, it provides nice guidelines for developing assessment simulations in many areas.

## **At what level should the standard be assessed?**

The indicators of standards vary in their complexity. In deciding how to create the artifact you will use for assessment and your scoring scheme for the assessment, you need to consider where you think students should be along some knowledge taxonomy. For instance, you might consider thinking about the levels of Bloom's taxonomy (knowledge, comprehension, application, analysis, synthesis, and evaluation). Yet another way of thinking about the level of performance necessary to succeed is to examine knowledge integration. One model for this is Biggs and Collis' (1982; Biggs, 1999) SOLO taxonomy. The SOLO taxonomy looks at not just whether a student knows

something, but how well he or she knows it. At lower levels in this taxonomy, student knowledge consists of disconnected facts. At higher levels, the student integrates knowledge and connects it with the big picture of how it fits with other related concepts, procedures, and so on. There are teacher education programs and higher education institutions that use the SOLO taxonomy. This taxonomy can not only help you to decide at what level of learning the standard requires, but may also help you in differentiating students who score at different levels on the State Scoring Rubric. For example, you could look at ways to differentiate students who knew concepts from those who not only knew the concepts, but showed some integration of those concepts.

Yet another taxonomy to consider is the one reported by Gagne, Briggs, & Wager (1992). Gagne et al. (1992) break down learning into several different domains, and differentiate several levels of intellectual skills and strategies. This taxonomy, like the others, can help you to think more clearly about what you need to assess and how you need to assess it.

### **How should I apply the rubric?**

Below is the generic rubric and its interpretation provided for us by the Alabama State Department of Education. In order to interpret the rubric, you need to translate the expectations into performance indicators that would reflect each of the four levels of performance.

#### **Rubric Interpretation**

The standard rubric used for evaluation and reporting the level of knowledge, ability, and/or skills is provided by the Alabama State Department of Education. It is a four point measure with detailed evidence descriptions provided on each evaluation and appears in the chart below. If you are constructing a rubric for an educational support field (e.g., library media, counseling), or if you are constructing a rubric for a field that does not have state certification requirements, you can easily modify this rubric by simply replacing the term “teaching professionals” with your particular field. If you are constructing a rubric in a noncertification program, then you can change initial level of certification to beginning level or some other adjective to identify the student as someone who would be just starting out in the field at the level of your program (e.g., masters, doctoral, etc.). Regardless, it is important to note that when interpreting the rubric, you need to pay attention to the descriptors of performance in relation to peers. As you can see, the main criterion for differentiating students is how well they did in comparison to peers at a similar level.

Level	Evidence
<b>4: Exceptional (4 pts.)</b>	The candidate demonstrates exceptional understanding and/or skill expected of teaching professionals at the initial level of certification. Knowledge conveyed and/or performance demonstrated regarding this standard places the candidate <b>at a level far beyond peers.</b>
<b>3: Proficient (3 pts.)</b>	The candidate demonstrates proficient understanding and/or skill expected of teaching professionals at the initial level of certification. Knowledge conveyed and/or performance demonstrated regarding this standard is consistent with that of <b>effective preservice teachers.</b>
<b>2: Basic (2 pts.)</b>	The candidate demonstrates basic understanding and/or skill expected of teaching professionals at the initial level of certification. Knowledge conveyed and/or performance demonstrated regarding this standard is consistent with preservice teachers' <b>initial understanding and/or performance in this area.</b>
<b>1: Unacceptable (1 pts.)</b>	The candidate <b>does not demonstrate minimal understanding</b> and/or skill expected of teaching professionals at the Class B level of certification. Knowledge conveyed and/or performance demonstrated regarding this standard is <b>unsatisfactory.</b>

Your job is to link your scoring with the rubric levels. For knowledge items, you might use criteria that involve indicators of the level of knowledge integration or application like those in the different knowledge taxonomies noted above (e.g., SOLO). You could also look at criteria from test results (a grade of 90% or higher on items related to the indicator is a four, 80% a three, 70% a two, and 69% and less a one. For items that involve performances, you will need to judge how the performance looks in comparison to prior students, in-service teachers, or other examples of expert or not-so-expert performance. Again, you should try to share your criteria with others so that you can get agreement concerning whether your scheme can be reliably implemented. If others cannot see the differences in performance, it is likely that you need to review your criteria. A good rule of thumb is that if a faculty colleague and an in-service expert teacher or other professional (depending on the program) finds the scaling plausible, you are probably on the correct path.

### **Developing a Matrix to Link the Standard, the Assessment, and the Rubric**

A useful way to help you to develop your assessment and provide some evidence of its content validity is to set up a matrix that links the artifact or activity with the generic rubric and the specific scoring criteria that need to be met. This is helpful in a variety of ways. First, it makes the assessment process clear to the students if you share that matrix with them. The criteria they need to meet will be transparent. Hence, they can work toward modeling expert practice (striving for a “four”). Second, it provides a record for your program and the college of the nature of your assessment. This will be helpful in program assessment and evaluation as well. It will provide evidence of content validation of the standards.

Below is a sample matrix for developing and reporting your scoring scheme. First there is a general description of the artifact, activity, or other task associated with the indicator. Next, the generic rubric is listed with a column for the specification you will use to link the scores to the rubric. Ask yourself, what are the traits associated with exceptional, proficient, basic, and unacceptable performance? Finally, in the last column, there is a place to justify why the performance you specify meets the rubric criteria (score of one, two, three, or four). If you feel comfortable that your score criteria could be explained in this matrix, then your assessment is likely to be well designed and defensible. A blank word version of this matrix is available from the Office of Assessment and Evaluation, if you want one to work with as you design or evaluate your assessments.

## **Worksheet for Developing a Specific Rubric for a Specific Standard or Indicator**

### **Indicator knowledge or Ability Statement:**

**(2)(c)5.(xii) Ability to Interpret and use reports from state assessments and results of other assessments to design both group and individual learning experiences.**

**Course: EPY 455**

### **Description of Activity or Artifact to be used to Meet knowledge or Ability Statement:**

Test score simulation. Students will be put in the scenario of a teacher who is getting ready for the start of the school year. He or she is given the ARMT, SAT, and CRT scores of the incoming students. The task for the teacher is to develop instructional plans for adapting instruction based on the patterns of scores.

## Scoring Rubric for Indicator

Level	Generic Standard	Specific Performance Traits associated with the task indicative of performance at this level.	Why traits are indicative of performance at this level
<b>4: Exceptional</b>	The candidate demonstrates exceptional understanding and/or skill expected of teaching professionals at the initial level of certification. Knowledge conveyed and/or performance demonstrated regarding this standard places the candidate <b>at a level far beyond peers.</b>	<p>Instructional Plan well specified</p> <p>Student is able to differentiate instruction and makes use of all the relevant information</p> <p>Student integrates the test information exceptionally well.</p> <p>Student proposes instructional strategies to improve performance that are creative and are very tightly linked to the test information.</p>	An exceptional student has all of the traits of a student who is proficient, and he or she is able to come up with exceptionally well designed instructional strategies based on test data. Creativity, integration, and innovativeness separate the exceptional student from the proficient one.
<b>3: Proficient</b>	The candidate demonstrates proficient understanding and/or skill expected of teaching professionals at the initial level of certification. Knowledge conveyed and/or performance demonstrated regarding this standard is consistent with that of <b>effective preservice teachers.</b>	<p>Instructional Plan well specified</p> <p>Student is able to differentiate instruction and makes use of all the relevant information (e.g., subscores, crt objectives that are indicative of potential problem areas)</p> <p>Student proposes conventional instructional strategies to improve performance</p>	A proficient student is able to use all the information available from test scores to determine areas of strength or weakness. He or she should be able to interpret the pattern of information and suggest relevant instructional strategies that should be effective.
<b>2: Basic</b>	The candidate demonstrates basic understanding and/or skill expected of teaching professionals at the initial level of certification. Knowledge conveyed and/or performance demonstrated regarding this standard is consistent with preservice teachers' <b>initial understanding and/or performance in this area.</b>	<p>Student shows a basic understanding of test scores</p> <p>Student is able to develop a plan to differentiate instruction, but does not make use of all the information available in the test scores</p> <p>Instructional plan is not well specified.</p>	At the very least, a teacher should be able to interpret test scores and make some instructional judgment about how to adapt instruction based on them. The limitation here is that the student does not develop a well specified plan of action based on the scores.
<b>1: Unacceptable</b>	The candidate <b>does not demonstrate minimal understanding</b> and/or skill expected of teaching professionals at the Class B level of certification. Knowledge conveyed and/or performance demonstrated regarding this standard is <b>unsatisfactory.</b>	<p>Student does not interpret scores correctly</p> <p>Instructional plan does not include differentiation of instruction plan based on test scores</p>	Students need to be able to understand test scores and use that information to improve instruction. If a student cannot interpret scores or come up with a plan, then they are not competent to be in the classroom.

## **How can the Reliability, Validity, and Fairness of Assessments be Evaluated?**

An instructor can help a great deal in content validation of the assessment. Getting agreement that the assessment is valid from colleagues who have some similar expertise, and doing a rational task analysis of what it is that you want students to do, can really help put the assessment process on a solid foundation. Also, it is important that instructors truly find ways to differentiate between score levels using valid criteria. Not everyone should score a four (nor a one)! Honest appraisal of the quality of student work along a well designed scale provides a valid assessment that can help the student and the program improve. Along with the validation process that you engage in on your own, we will be working as a college to check the validity of assessments in different areas. NCATE requires us to determine that our assessments are reliable, valid, and free of bias. As we collect assessments, we will be able to examine how they relate to student performance and success once students are out in the field. For example, we could link the results of your assessments to PEPE results or performance of students on related indicators. Additionally, we will work on sampling assessments so that we can check for interrater reliability and validity. While it is impossible to check all assessments, the Office of Assessment and Evaluation in the college will periodically check the interrater reliability and validity of some assessments. If you have special needs and wish to do a small scale interrater reliability or validity study, the Office of Assessment and Evaluation will help you set it up. Finally, the college will look at various demographic elements of success and failure rates to make sure that there does not appear to be bias in the results. You can also help to ensure fairness, by making your scoring scheme and performance expectations clear to the students. The better they understand what they have to do, the less likely it is for them to claim that you were not fair, or that they did not know what to expect.

Given the newness of this process for some of you, expect that you may want to make improvements in your assessments over time. This is to be expected, and you should report such changes and why you made them as part of the ongoing assessment process for your program and department.

## References

- Biggs, J. (1999, April). What the Student Does: teaching for enhanced learning. *Higher Education Research & Development*, 18(1), 57-75.
- Gagne, R., Briggs, L. & Wager, W. (1992). *Principles of Instructional Design (4th Ed.)*. Fort Worth, TX: HBJ College Publishers.
- Merrill, M. D. (2002). [First principles of instruction](#). *Educational Technology Research and Development*, 50(3), 43-59.
- Thorton, G. C., Mueller-Hanson, R. A. (2004). *Developing Organizational Simulations*. Mahweh, NJ: Lawrence Erlbaum and Associates.
- Wiggins, G. & McTighe, J. (2005). *Understanding by design: Expanded edition*. Alexandria, VA: ASCD.