CERIAS Tech Report 2005-149

Using outcomes-based assessment data to improve assessment and instruction: a case study

by Melissa Dark
Center for Education and Research
Information Assurance and Security
Purdue University, West Lafayette, IN 47907-2086

Using Outcomes-Based Assessment Data to Improve Assessment and Instruction: A Case Study

Steve Rigby
BYU-Idaho
Smith 418
Rexburg, ID 83460
208-496-1494
rigbys@byui.edu

Melissa Dark
Purdue University
Knoy Hall
West Lafayette, IN 47906
765-494-7661
dark@purdue.edu

ABSTRACT
Educators who have been through accreditation are well aware of the need for outcomes-based learning and assessment. However, there are misunderstandings about what outcomes based assessment is, and how it can improve teaching and learning. We understand that accreditation requirements can be a reason for adopting outcomes-based assessment, but our real goal is to convey to our readers how outcomes-based assessment can provide meaningful and useful feedback to the instructor regarding student achievement, assessment, and the quality of the instruction.

Categories and Subject Descriptors
K.3.2 [Computers and Education]: Computer and Information Science Education –Curriculum.

Keywords
IT education, curriculum, outcomes-based assessment.

1. INTRODUCTION
This paper is intended for those fairly new to the area of outcomes assessment, validity, and reliability, and provide an example of how an outcomes-based approach to teaching increases the overall effectiveness of a course.

'Objective-based' teaching is the process of identifying what students will be able to do after instruction. Objectives that specifically state student performance as a result of instruction are also referred to as 'outcomes'. Outcome statements usually start with verbs that reflect the performance(s) students are expected to demonstrate to indicate achievement of outcomes, e.g., identify, solve, list and select. Outcomes can be expressed for an entire degree, course, or lesson. The articulation of outcomes can be helpful for the instructor and student. Clearly stated outcomes give the instructor focus, guide the instructor to what resources should be used, and suggest performances to gauge student learning and instructional effectiveness. Clearly stated outcomes also provide students with a roadmap for instruction that they can use to focus and regulate their own learning [1].

Outcomes-based assessment is the process of developing assessments based on the outcomes that were created for instruction. Assessments can include quizzes, examinations, portfolios, performances and so on. We have found that many educators believe that if the scoring of the exam produces a bell shaped curve, then the exam is effective. However, forcing a normal distribution assumes 1) relatively large samples, and 2) is appropriate only when the goal of instruction is to compare students to each other as opposed to a criterion. In criterion-referenced assessment, the criteria for performance are derived directly from the outcomes, performance standards are clearly stated for each criterion, and the means for demonstrating achievement are explicit. As with all assessment, validity reliability, and practicality are important in developing and using criterion-referenced achievement tests effectively.

Validity refers to whether the assessment measures what we want to measure. One example of a threat to validity would be to test the student on material that does not relate to the outcomes that were taught. Another example of a threat to validity would be, given different outcomes that are considered of equal importance, to ask a disproportionate number of questions or to weight the items disproportionately.

Reliability refers to the quality of the measure in the sense that the assessment produces consistent or repeatable results. One example of a reliability issue for multiple choice and true/false questions would be how easily students can guess the correct answer without knowing the material. If students can easily guess answers, then the examination fails to reliably distinguish students who know the content from those who do not. Another example of a threat to reliability would be inconsistent grading of essay questions by changing the way the grader scores essay questions over time. Generally speaking, the more authentic and longer the assessment, the more valid and reliable it will be. Practicality refers to the reasonableness of the assessment; practicality is often at odds with validity and reliability [2]. For example, one of the best ways to measure if students have mastered an objective would be to provide a real-life example and then observe their solutions. However, this may not be practical or feasible given resource and time constraints, such as using multiple choice tests instead of projects and experiences for assessment or using a 50 item test, when a 200 item test would be more valid and reliable. There are always trade offs between a) validity and reliability and b) practicality, and the instructor must balance those given the nature of the decisions to be made using the assessment data. [1]
Outcomes-based assessment can go beyond providing feedback on student achievement. It can also be used to provide feedback on the effectiveness of instruction. Issues of validity, reliability

and practicality are still applicable, but now are based on an understanding of the underlying model of cognition for the instruction.

2. THE MODEL OF COGNITION

The model of cognition has three vertices as shown in figure 1 [3]. The first vertex is outcomes. As discussed earlier these are the desired performances as a result of instruction. The second vertex is instruction. This includes the selected instructional methods and techniques to effectively lead to the desired outcomes. The third vertex refers to assessment. This refers to both the combination of assessments (all examinations, quizzes, etc.) as well as the individual assessment items on each examination and quiz. When the model of cognition is used, the assessment data can be used to provide feedback on student achievement and also to provide feedback on the effectiveness of instruction.

Outcomes are placed at the bottom or foundation of the triangle because the articulation of outcomes should precede the development of instruction and the development of assessment. Instruction should be based on outcomes and so should assessment. Only when instruction and assessment are both based on outcomes can you use assessment data to provide feedback on the effectiveness of instruction.

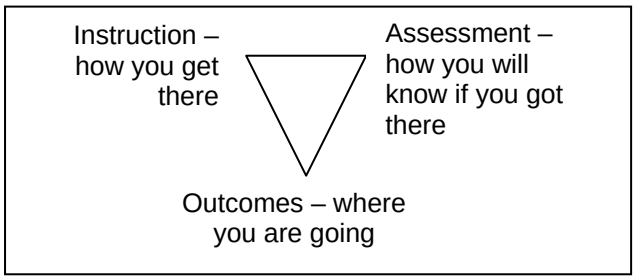

Figure 1. Model of cognition.

3. PUTTING THEORY INTO PRACTICE

Putting theory into practice, a model of cognition was used in an introductory Information Assurance course. Outcomes were developed for the course, instruction was based on the outcomes, and assessments developed were also based on the outcomes. We start with a discussion of how we attempted to ensure validity of the examination. We then move to a discussion of relevance of that to making decisions about individual achievement and how to use assessment data to make decisions about instruction.

3.1 Matching Exams Questions to Outcomes

To increase the validity of the exam, questions were developed based on the outcomes of the course. Table 1 shows all of the outcomes for the first half of the introductory Information Assurance course. Column three shows the relative importance of each outcome. The sum of the outcomes is obviously 100% as the outcomes are the totality of what you want students to achieve. Column four shows the number of questions on the examination that were used to test each outcome. Column five shows the points associated with each question and column six shows the points associated with each outcome. The concept that an exam may weight one outcome more than another is not widely understood. If each question on this examination were weighted equally, then a question that tested outcome #5 would be worth 1.2195% of the total (1 divided by 82) and all 13 questions that test outcome #5 would be worth 15.8536% of the total (13 divided by 82). However, by weighting the points per question (making each question for outcome #5 worth 1.846 points) these test questions are now worth 24% of the total. This should be a sequential process where the importance level is decided first and then how many questions there are for each objective is identified. Note that it is possible for there to be 100 questions on the examination if there are 100 distinct questions to be asked. It is not necessary to write 24 separate questions to test outcome #5 if 13 questions are sufficient to test the content area represented by this outcome.

Table 1.

Outcome #	Outcome	Importance	Number of Questions	Points per Question	Points per Outcome
1	Identify historical events in computer security.	3	3	1	3
2	Describe the various types of threats that exist for computers and networks, and costs associated to those threats.	5	5	1	5
3	Identify different avenues of attack.	6	3	2	6
4	Describe how a layered defense related to both physical and computer security.	4	4	1	4
5	Describe and define CIA and basic security concepts.	24	13	1.846	24
6	Describe the basic model of security.	3	3	1	3
7	Identify authentication protocols and their uses.	9	9	1	9
8	Identify good/poor security practices.	8	4	2	8
9	Describe why physical security education is important.	4	4	1	4
10	Identify the different algorithms and terminology of cryptography.	5	5	1	5
11	Identify the key components of PKI.	8	8	1	8
12	Describe encryption standards and protocols.	6	6	1	6
13	Identify network protocols, topologies, architectures, and administration.	15	15	1	15
	Totals	**100**	**82**		**100**

3.2 Point Distribution

After the students took the exam, a spreadsheet was created to record individual and group performance for each question. Table 2 below shows an example of this for questions 1, 2, 3, 81 and 82. The last two columns show how student performance differs using a weighted and non-weighted score. All of the students did as well or better using the weighted score over the non-weighted score. A majority of the students scored one to two percentage

points higher when weighted by objectives rather than by equal weight scoring. One student (not shown in this example) scored four percentage points better, which constitutes almost half of a letter grade. This could be the difference between a grade of A and B or a grade of B and C, etc. Needless to say, this becomes important to the students when applying for scholarships and or graduate schools and shows how content validity impacts the assessment of individual achievement.

Table 2.

Question	1	2	3	81	82	Weighted Score	Non-weighted Score
Outcome	3	2	5	13	13		
Student 1	2	1	1.846	1	1	92%	91%
Student 2	2	0	1.846	1	0	83%	83%
Student 3	2	1	1.846	1	1	90%	89%
Student 4	2	1	1.846	0	1	85%	84%
Student 5	2	1	1.846	1	1	80%	78%
Student 6	2	1	1.846	1	1	83%	83%
Student 7	2	1	0	1	1	73%	71%
Student 8	2	1	0	1	1	75%	74%

4. ASSESSMENT PRINCIPLES: ITEM DIFFICULTY AND ITEM DISCRIMINATION INDICES

As mentioned earlier, there are two main theories of assessing students' understanding of material after instruction; norm-referenced and criterion-referenced. The objective or norm-referenced assessment is to rank individual performance among students; this comparison among students is useful in situations like selection for entrance (the SAT for college admission would be an example). However, this type of assessment is not as useful for determining if the students understood the outcomes taught in the classroom. What is appropriate for this purpose is criterion-based assessment, which focuses both on how well the students understood the outcomes and identifies which students need remediation. With criterion-referenced assessment the goal is to improve instruction so that everyone can master the subject material. If it is determined that everyone answered a question correctly and the question is determined to be a valid, it will not be removed from the exam [1].

Two tools that are used for determining the validity of criterion-referenced assessment are the item difficulty index and the item discrimination index [2]. The purpose of both item difficulty and item discrimination analysis is to improve examinations by identifying ineffective test items and then rewriting them or deleting them. The item difficulty index is calculated by dividing the number of students who answered the question correctly by the total number of students [4]. For example, if 20 students took a test and half missed question 3, you would divide 10 by 20 resulting in an item difficulty level of .5. Questions with an item difficulty index less than .6 (more than forty percent missed the question) provides useful feedback for the instructor. When there are highly missed questions, the instructor can do some investigation and reflection to try to determine why the question was missed so frequently. Careful analysis of the examination question could reveal wording problems such as double negatives or multiple correct answers, which make the item potentially invalid. Questioning the validity of the question should always be the first step. However, if the examination question is deemed valid, then it is appropriate to move on to analysis of instruction. Reflection on the instruction could also be helpful when trying to assess why an item was missed. Was the topic covered too quickly? Was there enough time allowed for questions? This analysis can help improve future lectures so that the appropriate time is given to learn the objectives.

The item discrimination index can be used to see if the question is answered correctly more times by the students who scored above the median and was missed more frequently by those students who did worse than the median [3]. This is accomplished by dividing the students into two groups, those who scored above the median (called the "upper" group) and those who scored below the median (called the "lower" group). An item difficulty index is computed for the "upper" and "lower" half of the scores and the "upper" difficulty index is then subtracted from the "lower" index to compute the item discrimination index [3]. The resulting item discrimination index can range from -1 to 1. The interpretation of this index is that if everyone answered the question correctly the score would be 0. If everyone in the "upper" half answered correctly and everyone in the "lower" half missed the question, the item discrimination index would be 1. Conversely, if everyone is the "lower" half answered the item correctly and everyone in the "upper" group missed the item, then item discrimination would be -1. If the intent of the exam is to reward students who studied and prepared for the exam and penalize students who didn't, then a discrimination index score greater than 0 would suggest this was the case for that question.

The higher the discrimination index, the more dichotomy between the two groups. When the discrimination index falls below zero, this suggests that the "lower" half of the students did better on that question than the "upper" half. This should then prompt further analysis to determine why the students who performed poorly on the exam scored better on that question.

The discrimination index should not be used as the sole indicator for looking at the validity of exam questions. The difficulty index should also be used for analysis. An example of why the discrimination index should not be used as the sole indicator is when one question is missed by every student in the class. The item discrimination index for this question would be 0. If everyone in the class correctly answers a question the item discrimination index would also be 0. By looking at the item difficulty index along with the item discrimination index, a picture starts to come into view of the validity of the questions. Let's return to our example.

Table 3.

	Question														
Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1															
2															
3														X	X
4							X								
5							X								
6														X	
7													X		
8							X					X	X	X	
9														X	
10							X								
11		X					X								
12						X	X								
13				X			X								
14		X			X	X	X						X		
15			X	X	X		X					X		X	
16			X	X	X		X					X		X	
17						X	X	X					X		
18											X		X		
19							X								
20						X	X					X			
Item Difficulty Index (p)	1	.9	.9	.85	.85	.8	.35	.95	1	1	.95	.8	.75	.7	.95

One question with an item difficulty of 0 was missed by the entire class (not shown in the table 3). In reflecting over the course, it was decided that this topic had not been discussed. The question was an obscure recall question and should be dropped from the exam. Question number seven (seen in table 3) was missed by 75% of the class. This question reflected a topic that was mentioned briefly during class, and the difficulty index score indicates that further discussion on this topic should take place for comprehension. This type of feedback is helpful for improving the quality of instruction.

Although the questions with an item difficulty index from .6 to .9 represent a majority of the students answering correctly, there may be other validity concerns. Were these questions answered correctly because of the quality of instruction and the student's preparation level, or were these questions easily guessed and not reflective of the stated outcome. This is where combining the item discrimination index along with the item difficulty index can be useful.

4.1 Item Difficulty Example

For the information assurance course, the item difficulty index was calculated for each test question in the exam. Table 3 shows the partial results of the item difficulty calculation. The 'x' in the cells in Table 3 indicates the student who missed that question.

4.2 Item Discrimination Example

The item discrimination index can be used to help determine if the questions were missed by those who knew the material or those that did not. Table 4 shows the results of the item discrimination calculation. The "upper" half of the students are denoted by numbers 1 through 10, and the "lower" half in greyed rows 11 through 20.

Table 4.

	Question														
Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1															
2															
3														X	X
4							X								
5							X								
6														X	
7													X		
8							X					X	X	X	
9														X	
10							X								
11		X					X								
12						X	X								
13				X			X								
14		X			X	X	X						X		
15			X	X	X		X					X		X	
16			X	X	X		X					X		X	
17						X	X	X					X		
18											X		X		
19							X								
20						X	X					X			
Item Difficulty Index (p)	1	.9	.9	.85	.85	.8	.35	.95	1	1	.95	.8	.75	.7	.95
p (upper)	1	1	1	1	1	1	0.6	1	1	1	1	0.9	0.8	0.6	0.9
p (lower)	1	0.8	0.8	0.7	0.7	0.6	0.1	0.9	1	1	0.9	0.7	0.7	0.8	1
Item Discrimination Index	0	0.2	0.2	0.3	0.3	0.4	0.5	0.1	0	0	0.1	0.2	0.1	-0.2	-0.1

The item discrimination index range for the class was between -.3 and .6. Questions with a discrimination index from .1 to .6 fall into the desired rage of valid questions while those below zero need closer evaluation. An example of how the discrimination index can be helpful in determining the validity of a question is to look at question 14, which has a discrimination index of -.2. The question asked the student to identify a communication protocol and two of the answers could be argued to be correct by the students. The item discrimination index value of -.2 adds plausibility to this argument. Because of the ambiguity regarding which answer is correct, another choice should be added to replace the "distractor" answer that was thought of as correct. Common sense is also important to account for normal variation. Question number 15 has a negative discrimination index although only one student missed the question. This could be explained by an accidental selection by the student.

5. LESSONS LEARNED

After performing this analysis, a refining process took place where questions were either modified or eliminated based on the item difficulty and item discrimination index. It was decided that four questions would be removed from next semester's exam because of high difficulty indices. Another four questions with negative discrimination indices less than -.2 will be modified and/or revised with more instructional time dedicated to those topics. By this continuous process of post exam analysis, more was learned about the course, instruction, and the actual examination itself than was ever learned from student evaluations. Evaluations helped gain student perspectives of the overall course, while post-exam analysis has provided specific feedback for improving 1) the examination itself, as well as 2) instruction. Treasures of information were hidden in an unexpected location. Specific feedback on objectives, validity problems with questions, and the time required for students to master topics were gleaned from this analysis. Through this project, we came to understand the value of taking an outcomes-based approach to improve the practice of testing and teaching. Surprisingly, the time required to perform this analysis was minimal. Initial setup required an hour our two to create the tool to automate this process. The authors used Excel, but many other software packages could be used. Many University testing centers provide this information on all exams.

6. REFERENCES

[1] Smith., P., & Ragan, T. (1999). *Instructional design (2nd ed.).* John Wiley & Sons.

[2] Gronlund, N. (2005). *Assessment of student achievement (8th ed.).* Allyn & Bacon.

[3] National Research Council (2001). Knowing what students know: *The science and design of educational assessment.* Committee on the Foundations of Assessment. Pelligrino, J., Chudowsky, N., and Galser, R, editors. Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington DC: National Academy Press.

[4] Psychometrics. Available at http://www.psych.westminster.edu/psychometrics-ws/