Validity, Reliability and Item Bias

W. Blaine Dowler

May 28, 2011

1 Item Quality

An assessment tool is only effective if it is evaluating the skill or curricular outcome which it is designed to evaluate. For example, assume a curricular outcome is "student is familiar with milestone events in superhero comic books published in the 1960s" and one needs to test this outcome on an exam.[1] A question such as

1. Fill in the blank: Peter Parker is also known as ______________.
   (a) Batman
   (b) Hulk
   (c) Spider-Man
   (d) Superman

is a terrible way to measure. One can correctly answer this question having never read a superhero comic from the 1960s. In fact, there are millions of people alive today who would answer that question correctly, but who have never read a comic book in their lives. Effective assessment depends on identifying effective items and rejecting the ineffective items. The three primary measures which must be satisfied by an assessment item are *validity*, *reliability* and *bias*.

2 Validity

The first critical element to an effective assessment item is validity. An assessment item is considered valid if it measures the skill in question. These can

[1] I don't know which course would have this outcome, but I'd like to take it.

be extremely difficult to write when there is more than one way to arrive at the answer to a question. The above comic book example is one that obviously has multiple approaches to the answer: one could watch an adaptation of the character in other media, or read a comic book from another era which contains the same information.

For a more subtle example, assume you are trying to assess a mathematics skill outcome described as "student will be able to multiply a three digit number by a two digit number, with or without regrouping." The question $273 \times 87 = ?$ is more valid than $458 \times 11 = ?$. Let us examine why that is.

Both questions involve the multiplication of a three digit number by a two digit number, and regrouping is optional. They both appear to satisfy the criteria of the skill outcome. However, in the second problem, the two digit number is 11, which is a relatively low two digit number. There are students who will solve this by adding 11 instances of 458. Thus, it is never actually determined whether or not the student can do the multiplication; the student who uses explicit repeated addition does so because he or she is unsure about how to correctly multiply by two digit numbers. The importance of using the "place holder" 0 in the second line is not understood. However, the first question involves multiplying by 87. This will ultimately be more valid, as there are far fewer students who will add 87 instances of a three digit number correctly. In this case, the validity differences between the two items is not as significant as the comic book question, but there is a definite difference.

Careful item creation can improve the validity of items by simply choosing items in which alternative methods are either difficult to implement or entirely impossible. However, the validity of the question can still be in jeopardy. If the item is a multiple choice or matching item, then it can be invalidated by a lack of plausible distracters. In other words, if the possible wrong answers are clearly wrong, the student can deduce the correct answer with no understanding of the skill. Let's return to the comic book outcome to illustrate this. Look at the following item:

1. Fill in the blank: Doctor Octopus first appeared in ______________.
   (a) Amazing Spider-Man #3
   (b) Archie #4
   (c) Batman #7
   (d) drag

The prompt is one that would seem to require a knowledge of superhero comic milestones of the 1960s, which is what we want. However, three out of four possible responses are clearly wrong. Even without the success of the recent movie series, it isn't hard to imagine that anyone taking a course with an outcome such as this would know that Doctor Octopus is most strongly associated with Spider-Man, so the correct answer stands out from the others like a beacon. A better version would be

1. Fill in the blank: Doctor Octopus first appeared in _______________.
   (a) Amazing Spider-Man #2
   (b) Amazing Spider-Man #3
   (c) Amazing Spider-Man #4
   (d) Amazing Spider-Man #5

In this case, students would either need to know the correct answer, or know the content of issues 2, 4 and 5 well enough to know that those could not possibly be correct.

The easiest way for an experienced teacher to improve the validity of the plausible distracters in multiple choice items is to determine the most common mistakes, and use those responses to complete the incorrect options. This is particularly important in math and mathematical science courses, in which students may realize their calculated response is incorrect when it is not found in the list of responses. For this reason, math and mathematical science teachers should seriously consider including "none of the above" as an option on every calculational question, and using it often enough to make it a plausible option. Note that overuse of this option can also reduce the validity of a question: students who have so little understanding of a topic that they fail to produce any of the numbers listed as common errors will correctly respond "none of the above" when they do not truly understand the skill.

Of course, it is entirely possible that the teacher is unable to recognize that an invalid item is invalid. As a human being, it is a natural psychology: if one is trying to write an item that can be solved by method A, it may not cross one's mind that another approach is possible, particularly when method A is the most efficient approach. One would need to apply the item analysis techniques of lesson six to determine whether the items truly perform as anticipated on the assessments.

3 Reliability

The reliability of an assessment item is similarly important. Reliability speaks to reproducibility. A reliable assessment item is one which will produce the correct response from students of sufficient ability over multiple applications. An item can be valid but not reliable, or vice versa.

Imagine a metaphorical dart board. If a student answers a question correctly, the dart hits the bulls eye. If a student has a misconception, the dart will land outside the bulls eye. If the item is valid, understanding the skill in question will be the only way to hit the bulls eye. If the item is reliable, the darts a given student throws will always land near each other.

Now, imagine an item which is valid, but not reliable. The darts will be scattered evenly across the board. The average position will still be the bulls eye, but the darts will be spread out across the board (and possibly the wall it is hanging from.) An example of a question of this type would be measuring the curricular outcome "student can multiply numbers with several digits with regrouping" by asking "what is $72,695,402,394 \times 42,394,634,630$?" It is virtually impossible to answer that question without understanding the skill, but the number of individual arithmetic operations involved make small mistakes very likely, so students who do have a fair understanding of the skill could quite possibly get it wrong. It is not a particularly reliable assessment item.

Conversely, imagine an item which is reliable, but not valid. Results are reproducible, but not accurate. The darts will still be bunched together tightly, but not near the bulls eye. One common example of this would be the "does your child have ADHD?" questionnaires that are often distributed. The same parent will likely give the same response each time he or she is asked to respond to the questionnaire, but the questionnaires are not accurate diagnostic tools in many cases.[2]

A reliable and valid question will see students with similar averages and abilities responding correctly and incorrectly with similar frequencies. If the plausible distracters are sufficiently well crafted from common mistakes, one may even find that students with similar ability levels below the difficulty of the question tend to choose the same incorrect response.

[2] As is probably quite obvious, this is a pet peeve of the author. Varying reports put the number of students misdiagnosed with ADD and/or ADHD at anywhere from 80% - 90% of the students who have been given the label. True ADD and/or ADHD results from an inadequate blood flow to a certain lobe of the brain, and the brain craves more sensory input than it receives. As a result, student attention drifts to look for new stimuli, or in the case of ADHD, the student moves to generate the input cravings that the environment alone cannot satisfy. The true conditions are unmistakeable before the student reaches school age, impact all subjects which do not involve unusual amounts of physical activity, and cannot be reliably diagnosed without a medical professional and possibly a CAT scan. A student who exhibits ADD symptoms in a single subject only is more likely to be struggling in that subject, to the point where maintaining his or her attention on the subject for a full class is frustratingly difficult, and the student's attention wanders to take a mental break or to look for comprehensible stimulus. It is akin to watching a foreign film without subtitles; if you cannot understand all of the content relatively easily, your attention *will* drift.

4 Bias

Bias occurs in almost all assessments. It occurs when student performance on assessment items varies between two groups of students with equal ability levels. As such, bias skews both the validity and reliability of assessment items, so we need to make deliberate efforts to reduce bias as much as possible.

The most obvious source of bias is language. A student who has consistently gotten marks around 75% in English language science courses from grades K-8 will likely continue to do so. Now, imagine a student who consistently scored around 95% in Spanish language science courses in the same K-8 range. Now this Spanish student (who has been learning English for, say, 6 months) moves to the same school as the first student, and starts taking the same science course in English. The student who used to get 95% will likely see a sudden and severe drop in grades due to this language difference, which is a form of bias.

Now, if a course is in English, it is in English. There is nothing the teacher can do about that. However, the teacher *can* control the level and complexity of English used to teach the course. If the grade nine concepts can be effectively explained using grade six grammar and vocabulary, then using the grade six grammar and vocabulary will reach a larger number of students, despite their difficulties with the language. If assessment items are also written at this level, bias is reduced on the exams.[3]

There are far more subtle sources of bias. The most common sources are cultural, which become particularly pronounced in an ethnically diverse classroom. There are gender biases as well, primarily due to cultural differences. If you walk into a typical North American seventh grade classroom and ask students what C4 is, you will typically find that the male students are far more likely to know the answer than the female students.[4] If you were to ask the same question in a war torn nation, you will likely find the bias disappears, because exposure to the information is far more widespread. This can be difficult to do while still assessing students at all levels of Bloom's Taxonomy. To achieve the highest levels, one must ask students to apply knowledge, skills and concepts to personal experiences, which means the teacher must be careful to select personal experiences that would be shared by all students. This is extremely difficult to do in a diverse classroom.

[3] Obviously, this doesn't work in all areas or in all subjects. It's impossible to effectively measure grade nine level grammar using grade six level grammar, for example.

[4] C4 is a type of plastic explosive.

5 Upcoming Lessons

The concepts of validity, reliability and bias will be explored in more detail in lesson six, at which point we will be equipped with the mathematical tools necessary to perform detailed analyses of specific test items.