Broad Types of Assessment

W. Blaine Dowler

May 28, 2011

1 Classifying Assessments

In the previous lesson, we established that assessments are performed for different people and with different goals. The specifics of some of these differences are detailed in this lesson.

2 Education Stage Progressions

One way to divide assessments is by the stage in the education process at which they are administered. These are broadly grouped into *formative* assessments, in which students are still forming their understanding of the skill in question, and *summative* assessments, in which students are expected to have already learned the material, and are simply being assessed on where they stand with it.

2.1 Formative Assessment

Formative assessments can be marked in a variety of ways, and their use is a point of debate. The most common formative assessment is homework.

When a new topic is taught, the knowledge and comprehension levels of Bloom's Taxonomy are expected to be reached from the lesson alone. These levels are verified, and the application level is reached, through daily homework

assignments.[1] Homework is a hot topic of debate right now, with a few questions surrounding it.

1. *Should homework contribute to the report card grade?* In an ideal world, homework would not contribute to final report card grades. A final report card grade should reflect the student's ability to apply and demonstrate understanding of the skills and curricular outcomes at the end of the course. Homework is a formative assessment; this is the opportunity students have to test their initial understanding, gain feedback from the teacher, and identify and correct their misunderstandings. As such, the understanding students have when the homework is first completed is expected to be less than the understanding they have when the course ends.

   Sadly, we do not live in an ideal world. If homework does not contribute explicitly to the report card grades, there is a natural human tendency among the students not to do the homework. Without doing the homework, the typical student will not pass the second level of Bloom's Taxonomy, and the skill will not be retained. The practice required to succeed at the end of the course is not obtained. Using homework to contribute to the report card grades reduces this problem to a degree, but then the report card grade will not necessarily reflect the student's understanding at the end of the course.

   There is no clear cut answer to this question, though the current trend is to record homework results on a completion basis rather than marks, which may or may not be worth a small percentage of the final report card grade.

2. *How much homework is the right amount?* Traditionally, the value of practice has been well known, but the application has been through assigning large amounts of homework on a particular topic the day that topic is taught in class, and then never assigning homework on that topic again.

   This doesn't work very well.

   Research has shown that doing more than 5-10 problems of a particular type in a 24 hour period is not beneficial to the student. Thus, giving out 10 math questions today will be about as helpful to the students as giving 30 or 50 today. Now, that's not to say doing 30 questions total isn't useful, but doing, say, six questions for homework every night *for five consecutive*

[1] For reasons the author has never understood, homework rarely exceeds the first four levels of Bloom's Taxonomy, despite the fact that unit and year end exams typically do. It is the author's belief that all six levels should be challenged on the homework as well, so that students are better prepared when they are challenged with these levels on exams. This is particularly confusing in cases in which the textbooks provide questions at all six levels of the taxonomy, and homework assigned from the textbooks is deliberately chosen to omit the higher level questions because "they are long."

nights will be far more helpful when trying to ensure retention. Spreading out the topics like this not only encourages retention, it provides review questions which students will find less time consuming, and it helps them make connections between consecutively taught and related topics.

3. *Should homework be taken home?* If a student does have a misconception or misunderstanding after instruction, doing homework can turn that misconception into a habit. Once a habit is formed, it can be very difficult to break. Furthermore, with a dishonest student, there is no longer any guarantee that the student taking credit for the homework was the individual who actually did the homework. This is particularly true with students who are struggling to levels at which they are frustrated. Most parents don't like to see their children in emotionally stressful situations, and some will do the homework for the child instead of watching the student suffer.[2]

Research is indicating that the answers to these questions require breaks from tradition. This can cause a lot of friction between teachers and parents, particularly when the parents in question are those who were successful in the traditional school system. They need to be convinced that the changes are truly in the best interests of their children. Nothing will raise a parent's ire more than endangering his or her child's future, so changes in this area have been slow to come to avoid these confrontations.

2.2 Summative Assessment

Summative assessments are the ones used at the end of a unit or course, when the instruction on skills assessed has been completed. In an ideal world, in which emotion has no impact on student performance, these would be the only marks used for the final report card grades, simply because these are the marks that indicate assimilation of information after instruction.

Again, the real world is not ideal.

The problem here is test anxiety. Every individual has an optimum stress level. If a person feels too little stress, that person's performance is poor. As stress builds, the individual's focus improves, and performance improves. If the stress level is too high, performance crashes hard. When summative assessments are worth larger and larger proportions of a student's mark, the stress level in that student increases, particularly when that student has future goals that depend on the marks in that course. The student stress goes beyond the

[2]The real solution to this situation is to go back and fill in the foundation from previous years, but I have yet to see a public school system that does this effectively.

ideal, and the performance crashes. Students who suffer from this test anxiety will then have summative assessment marks that are less indicative of their understanding than their formative assessment marks.

It should be noted that true test anxiety will impact multiple courses, and often impacts all of them. If a student hasn't reached the highest levels of Bloom's Taxonomy in a course, then that student will have poor retention in that particular course. The student may also have higher homework marks than exams in this case, particularly if additional support is provided for homework, because the student will be "going through the motions" on the homework without understanding why those motions are correct. Information will not be retained for the exam, and student performance drops once more. If exam marks are lower than homework for one student in one class, this cause is more likely than test anxiety.

3 Measurement Type Divisions

Final grades on an assessment or in a course can be reported in two primary ways, depending upon the intended goals for the reporting. The two main categories are norm referenced and criterion referenced reporting.

3.1 Norm Referenced

Norm referenced reporting is designed to report on a student's performance *relative to* his or her classmates. The report does not indicate the proportion of the course that the student succeeded with.

The advantages to this method of reporting apply almost exclusively to prospective employers and educational institutions. Those entities are primarily concerned with identifying the "best of the best," and compare the performance of one student to a group of peers. Those peers may have been chosen by age or grade, and they may or may not have met. These assessments are also frequently machine scored, which makes marking them remarkably efficient.

The disadvantages cannot be ignored. First of all, a student can see the results of this assessment, and know that he or she did "average." What does that mean? What was the average? Who was in the norming group? Which specific skills must a student improve in to increase this performance?

The basic methodologies for norm referenced assessments are pretty universal, and must be understood in order to correctly interpret the results of

such an assessment. In most cases, no single school can provide a population base large enough to generate reliable statistics for these assessments, so they draw students from multiple schools in multiple districts. The assessment is administered to these students before it is released to the public.

The results of all of these students are then compiled, and collated by total score. Individual assessment items are evaluated for quality using criteria that will be described in detail in later lessons, bad items are rejected, and student scores on only those assessment items kept are then compiled and compared. Performance of students on this assessment tool in later years is determined by determining where they fit relative to the members of this original "norming" group.

This is not a particularly accurate system in many cases. Often, in order to use enough students to build the reference results, the test is administered to a random sampling of schools in multiple regions using different curricula. In that case, a student's performance relative to the entire group may differ from his or her performance within the local regional group. As a result, a student's reported performance may not align with the local standards.

Things get worse when looking at improvement. Effective assessment items are difficult to write. As a result, students either cannot see or cannot keep test papers after they are written, which makes it difficult to analyze errors and improve future performance.[3]

The final problem is one that, sadly, seems rather pervasive in the field of education. When building a norm referenced assessment tool, one works by comparing the data obtained to one or more models. Unfortunately, it appears to be a common practice to decide on a model in advance, and reject data that doesn't fit the model rather than adjust the model to fit the data. The models used are good enough to apply to the majority of the population, but as student performance drifts further and further from the average performance of a student, the model fits less and less well. Ultimately, the results of the population's highest and lowest performers tend to be rejected and omitted from assessment results used for comparison.

For example, think about IQ tests. In North America, IQ tests try to measure a person's overall intelligence by modeling test performance on a bell curve (also known as a normal or Gaussian curve) with an average of 100 and a stan-

[3]Note that, in many areas, such school policies may be illegal. Technically speaking, the student is the author of any document he or she produces. In most regions, this means the schools cannot legally deny a student request to receive a copy of any work he or she has handed in. Note further that this only applies to the actual student output itself, and not to the assessment. If the student is told not to write in the test booklet, and to answer on a machine scored multiple choice form, then the school is only required to provide a copy of the multiple choice form, and *not* the assessment questions that provide a meaningful context for that form.

dard deviation of 15.[4] In short, approximately two thirds of the population will score between 85 and 115 on a North American IQ test. Even if we correctly compensate for language differences that occur (i.e. very smart people who only speak Spanish will get terrible scores on an English-based IQ test) and other regional differences, we have problems with the model that cannot be overcome. A bell curve is completely symmetrical about the average. This means that we would have just as many people scoring 80 points below average as there are scoring 80 points above average. This isn't the case. The number of people with tested IQ scores above 200 is far greater than those with negative scores. This skew distorts extreme results. Most standardized IQ scores will "report" scores as high as 160. However, an individual with an IQ of 145 could easily get the 160 score on most tests, as the average difficulty of the questions is around the 100 IQ mark. Once a person performs far enough above average, the odds of a perfect score increase.

Every norm referenced assessment also has two ambiguous cases. A student who gets every question correct or every question incorrect no longer fits on the scale. A student with a perfect score on an IQ test which accurately measured from 70 to 130 could have an IQ anywhere from 135 to ∞. Similarly, a student who gets every response incorrect cannot be accurately measured (and probably doesn't understand the instructions or know the language the test is written in.)

Norm referenced reporting on larger scales, such as University classes, leads to a new set of problems. If grades are normed and averaged on a class by class basis, then it becomes difficult or impossible to evaluate the performance of individual teachers or a complete course curriculum. The average mark and spread of grades in the class section with the most effective instructor looks identical to the average mark and spread of grades in the class section with the least effective instructor, simply because they are forced to conform to the same scale.

Research has shown that, when the goal of assessment is to communicate with students and inform them of their strengths and weaknesses, norm referenced reporting is the single most *ineffective* reporting method developed in recorded history.

3.2 Criterion Referenced

Criterion referenced assessments do not depend on mathematical models or on the performance of large groups of people. Instead, they measure a student's performance in direct comparison to the curriculum of skills being assessed.

[4] European IQ tests typically use the 100 average, but often use a standard deviation of 10 instead.

Detailed criterion referenced assessment is the type of assessment that benefits the students the most.

Criterion referenced assessment is well suited to reporting broad pictures of a student's performance in a complete course, or to reporting how each student is performing with individual skills. Less detailed versions include letter grades and percentages; students will know how much of the unit or course they need to improve, but not which specific aspects need to be reviewed. More detailed versions which give a skill by skill breakdown, or which include personalized teacher comments, tend to give the students the best possible chances to improve future performance by directing them to the exact skills which need to be improved. The drawback is that it requires a considerably higher amount of time to mark. Given the current trends in North American class sizes, teachers simply cannot provide this level of information in the time available without obtaining or developing a computerized system that helps them increase efficiency.

4 Upcoming Lessons

In lesson three, we begin to discuss validity, reliability and bias of assessment items, which are the three key concepts used to evaluate the quality of the assessment tools themselves. Lessons four and five will discuss methods of norm and criterion referenced reporting in detail, and lessons six through nine will cover increasingly mathematical methods to build assessment tools which can achieve the best of both worlds.