Single Classroom Analysis

W. Blaine Dowler

May 28, 2011

1 Preface

The mathematics involved in this lesson are optional, but still minimal. An adept fifth grader could likely handle the math in this lesson. (The same is not true of the math in later lessons.) All lessons involving math will be handled the same way. An initial conceptual overview covers a completely non-mathematical description of the relevant topics, which will provide all background needed to follow the conceptual discussions in later lessons. After this conceptual overview is complete, an optional section with the full mathematical glory of the topic follows.

2 Evaluating Assessment Item Quality - Conceptual Overview

Teachers of a small number of classrooms do not have immediate access to the advanced analysis techniques available to large scale assessments due to the low student populations. However, there are a number of highly useful tools available to teachers on these scales which can be implemented to good effect.

It is also important to note that one should always analyze multiple assessment items for a single curricular outcome. If one only has a single assessment item and the class performs poorly, one is unable to determine if the problem is with the assessment item itself or with the lesson plan for the class in which that curricular outcome was taught.

The three key analyses which must be performed evaluate the difficulty, discrimination and bias of various assessment items.

2.1 Difficulty

The difficulty of an assessment item is the most intuitive concept. How difficult is the item for the students in question? This is also the analysis which benefits most from having multiple assessment items for a single curricular outcome. This also speaks to the validity of an assessment item.[1] If a series of items relating to the same curricular outcome are valid, they will have comparable difficulty values. If an analysis of the assessment items reveals inconsistent difficulty values, then one or more of the assessment items has a problem.

The difficulty of a concept is inherent to the concept itself. The overall class understanding of that concept will depend primarily on the difficulty of the concept and the effectiveness of the instruction that introduced the concept. Thus, if a class is given three or more different assessment items for this curricular outcome and produce two (or more) distinct levels of performance on that outcome as measured by those items, then the validity of the items is in question. Imagine at first that only one of the items is out of line with the difficulty of the others. If it is more difficult than the rest, it is likely that the item is poorly phrased or presented, so that the students are unable to recognize the item as an application of that particular curricular outcome. If it is less difficult than the rest, one must look carefully to see if students are solving the problem by an alternative means, reducing the validity of the item. With only two items, it can be difficult to determine which is anomalous. If there is only one item, comparison becomes impossible. Note that items being compared need not be on the same assessment. A quiz, a unit exam, and a final exam can all be compared, although interpretation can get harder. Should the students perform poorly on a quiz, it is likely the material will be reviewed and retaught, making later assessment items appear less difficult because instruction has improved. (In this case, strongly consider rejecting the original quiz from the final report card grades.) Should the performance on later items decrease, then it is likely that students did not have sufficient time to fully assimilate the information and move through all levels of Bloom's Taxonomy, in which case the information about the skill was not retained by the students. Future courses should adjust the time spent on relative topics.

The difficulty of valid, reliable items should be maximized through improved instruction. There is no reason teachers shouldn't aim for the highest possible average from their students, provided that high average is an *accurate* representation of the understanding students have demonstrated for the curricular outcomes assessed.

[1] Validity was first discussed in section two lesson three.

2.2 Discrimination

The discrimination of an assessment item is indicative of the item's reliability.[2] A reliable item will discriminate between the highest and lowest performers in a class. Top performers will answer the item correctly more often than low performers.

The concept behind calculating discrimination is simple: one the entire assessment has been graded (and not just the single assessment item in question), sort the entire class by their grades. Choose two equally sized divisions of students, where one division has the highest performers, and the other has the lowest performers. Compare the two groups. If the group of high performers performed better on the individual item than the low performers, the item is functioning properly. If the average performance in the two groups is comparable, then the item is not a reliable item, and should likely be rejected from any final calculation of grades. If the low performing group outperforms the top performers on the item, there's something terribly wrong. This typically happens on multiple choice items with a mistake on the answer key. If the answer key is correct, the item must be rejected: it is *not* a reliable item. Generally speaking, the most reliable items are the most discriminating items.

2.3 Bias

An assessment item that exhibits bias will be both invalid and unreliable. A bias in an assessment item means students can be lumped into groups which have different levels of performance despite comparable skill levels.

For example, imagine a class in which the average grades from female and male students are comparable.[3] To identify gender bias on a particular assessment item, perform the same steps done when analyzing discrimination, with the exception of choosing the groups. Instead of grouping students by grade, group them by the criteria which defines the bias. In this case, use one group with all of the female students, and another group of all male students.[4] If there is a significant difference in the performance of these two groups, then one needs to examine the structure of the item; it somehow relates to information from the local culture which has a gender bias within.

The next most obvious groups which may have bias are those who are learning in a second language.[5] This bias can be hard to eliminate; all one can do is

[2] For a refresher on reliability, see section three of lesson three.

[3] This should be every class with statistically significant populations of both genders.

[4] Androgenous students can be safely ignored in this analysis.

[5] This assumes that most of the class is learning in their first language. Roles are reversed

create assessment items using the simplest language possible. If the purpose of the class is to teach the language, such as English class, this may not be possible at all.

Finally, one needs to pay particular attention to opinion related assessment items. One should always check for bias comparing the papers written by students who support the same opinion as the teacher to those written by those on the other side of the argument. If there is a skew towards those who agree with the teacher, then the teacher may not be grading the assessment items based solely on the information presented by the student presentations.

3 Mathematical Overview

The mathematics in this lesson, as mentioned earlier, could likely be handled by an elementary school student. All readers are encouraged to read on. If the math is overwhelming, but the conceptual overview is understood, the reader may abandon the rest of this lesson without fear of being lost in the conceptual discussions still to come.

3.1 Difficulty

The mathematical definition of difficulty is simple, but counterintuitive. The higher the difficulty value an item has, the *easier* the item; the most difficult items have the lowest difficulty score.

The difficulty of an assessment item is the average score students achieved on that item. Mathematically speaking, this can be computed most easily as

$$\text{Difficulty} = \frac{\text{Total marks earned by class}}{(\text{Student population}) \times (\text{Maximum value of item})}$$

If the item is worth a single point, as with most machine scored items, this reduces to

$$\text{Difficulty} = \frac{\text{Number of students who answered correctly}}{\text{Number of students who wrote the assessment}}$$

So, in a class of 35 students, if 30 students answer a single-point multiple choice question correctly, the difficulty is $\frac{30}{35} \approx 0.86$. A difficulty of 0 indicates all students got the item wrong, and a difficulty of 1 indicates that all students got the maximum possible mark on the item.

in immersion classrooms.

3.2 Discrimination

The discrimination calculation begins by separating the student population by grades and performance. When choosing the groups of highest and lowest performers, try to take the top and bottom quarters of the class population. Also try to have at least ten students in each group; you may need to take the top and bottom thirds of the class to achieve this. Make sure you have the same number of students in each group. Try not to simply divide the class in half, as there will be too many borderline students in each group who will skew the results towards neutral discrimination.

With the groups selected, the discrimination becomes

$$\text{Discrimination} = (\text{Average mark of high performers}) - (\text{Average mark of low performers})$$

In the case of assessment items worth a single mark (such as most machine scored questions) and equal sized groups, this can be written

$$\text{Discrimination} = \frac{(\text{Correct high performers}) - (\text{Correct low performers})}{\text{Number of students per group}}$$

Perfectly discriminating questions have discrimination of 1, although values of 0.4 to 0.6 are typically the highest one expects in practice. (It is difficult to exceed these values without having extremely difficult questions, though one should aim for the highest discriminating non-biased questions possible.) A question with 0 discrimination doesn't discriminate: the two populations had equal performance on the assessment item. A question with negative discrimination saw better performance from the low performers, which is a serious problem with the question, and typically indicates an incorrect answer key.

3.3 Bias

The mathematics used to check for bias on a rudimentary level are virtually identical to those used to calculate discrimination. The only difference is the selection of the groups. One wants to group the entire class by the trait which may be biased. As this is unlikely to have equally sized groups, one should use the first formula, with the average marks within groups, to equalize the comparison. The discrimination when checking for bias should be at or near 0.

4 Upcoming Lessons

The next three lessons focus on item analysis of large scale assessments. The specific content within is not typically covered at the undergraduate level, but the shift towards large scale assessments built on this infrastructure is so clear that today's teachers, parents and students need to have at least a rudimentary understanding of the underlying concepts.