**Improving Your Questions Using Item Analysis**

Most test scanners and test software automatically create question and test statistics. These are most commonly known as item analysis. It is a numerical process that analyzes each question, or “item”, and relates how well students did on the question versus their overall test score. It is most often used to identify questions that are not working properly. For example, a difficult question that high-achieving students consistently get wrong and low-achieving students consistently get right might point at a question construction error. Item analysis lets you find these problems.

In the table below is a typical item analysis chart. It consists of two major portions: an analysis section that shows the success frequency, and a tally. The success frequency displays the percentages (or sometimes the raw counts) of students with the correct answer when considering the entire class (Whole Group), the top quartile of the class (Upper 25%), and the bottom quartile of the class(Lower 25%). The tally frequency section indicates the number (or sometimes percentage) of all students selecting each option. The percentage correct across different subsets of students can be used to determine the discrimination of the question.

|  |  |  |
| --- | --- | --- |
|  | Correct | Tally |
| Q | WholeGroup | Upper 25% | Lower 25% | Discrim | A | B | C | D | E |
| 1 | 81 | 97 | 56 | .46 | 8 | 6 | 79 | 1 | 3 |
| 2 | 55 | 53 | 59 | -.06 | 1 | 27 | 9 | 54 | 7 |
| 3 | 99 | 100 | 96 | .18 | 0 | 96 | 0 | 1 | 1 |
| 4 | 100 | 100 | 100 | .00 | 98 | 0 | 0 | 0 | 0 |
| 5 | 89 | 87 | 74 | .14 | 5 | 4 | 2 | 87 | 0 |
| 6 | 89 | 87 | 74 | .20 | 18 | 42 | 9 | 13 | 16 |

Table - Typical Item Analysis Chart

There are two useful ways to use the item analysis in your TBL course:

* In real time, after scanning the iRAT test sheets during the RAP process, you can use the tally frequency portion of the item analysis to identify questions with which students had difficulty. Understanding where students are having difficulty can help the teacher provide appropriate and timely clarification on any troublesome concepts after the tRAT.
* In off-season, the item analysis results can be used to revise and improve your questions. By reviewing the discrimination values of a particular question, you can understand how the question performed in the test. By reviewing the Tally Report frequency with which each option was chosen, we can identify options that may have never been chosen and need to be improved or made more attractive so some students might choose them.

**Using Discrimination Value**

Discrimination Index (“Discrim” in Table above) represents the ability for a question to differentiate between students with a good overall understanding of the test topics (i.e. those with a high overall test score) and those with a weaker overall understanding (i.e. those with lower overall test scores). The Discrimination Index value can range from -1.0 to +1.0. A large positive value indicates the performance on that individual question is strongly correlated to the overall test score. As the discrimination approaches zero, it indicates no relationship between question and overall performance, and as it becomes negative, it indicates students who didn’t prepare are more often answering the question correctly than students who did prepare! This usually indicative of an issue with the question wording or topic. The aim is to create questions that have a higher, positive Discrimination Index, showing that we are discriminating between the well-prepared students who understand the material, and those who did not prepare or do not understand the material.

In the example below, Question 2 has a low discrimination value, giving us the information that students who didn’t prepare performed better on this question than students who did prepare. This question probably should be revised.

|  |  |  |
| --- | --- | --- |
|  | Correct | Tally |
| Q | Whole Group | Upper 25% | Lower 25% | Discrim | A | B | C | D | E |
| 2 | 55 | 53 | 59 | -.06 | 1 | 27 | 9 | 54 | 7 |

In the example below, Question 1 has a higher discrimination value. This indicates students who prepared performed better than students who did not prepare. This question probably needs no revision.

|  |  |  |
| --- | --- | --- |
|  | Correct | Tally |
| Q | Whole Group | Upper 25% | Lower 25% | Discrim | A | B | C | D | E |
| 1 | 81 | 97 | 56 | .46 | 8 | 6 | 79 | 1 | 3 |

Finally, in the example below, all 98 students correctly selected answer A in Question 4. The question has a discrimination value of 0, indicating it does not discriminate between those with strong and weak overall test performance. This question is a candidate for revision on the basis that it is likely too easy. It would be especially important to revise this question if you suspect some students were not well prepared for the test, and yet they still were able to answer the question correctly.

|  |  |  |
| --- | --- | --- |
|  | Correct | Tally |
| Q | Whole Group | Upper 25% | Lower 25% | Discrim | A | B | C | D | E |
| 4 | 100 | 100 | 100 | .00 | 98 | 0 | 0 | 0 | 0 |

Typically, when questions discriminate poorly, the writers will re-examine the question and how it has performed to determine if revision is required. There isn’t a discrimination value threshold that clearly delineates a good question from a bad one. Discrimination values give us useful inferences to determine if the question is doing what we want it to.

**Using Tally Reports**

By examining the tally section, teachers can identify answer options that were never selected. This might point to the need to re-write some options to make them more attractive, forcing students to more accurately discriminate between options. One extreme case is the example above, where no student selected options B through E.

**Using** *p***-values**

Many test scanners will also provide question *p*-values. The *p*-value represents the difficulty of the question. For example, a *p*-value of .6 means 60% of students selected the correct answer. It is recommended that a test contain a range of *p*-values from .25 to .75. An easy question would have a *p*-value approaching .25 and a difficult question would have a *p*-value approaching .75. As previously mentioned, a prototypical RAP mix is 30% simple recall/remembering (did you do the readings?), 30% understanding (did you understand?), and finally 40% simple application, often in the form of “which concept applies to this situation” (are you ready to use what you know?). Therefore, on a ten-question RAT, one possibility is to have four very difficult questions, three medium-difficulty questions, and three easy questions.