Analyzing Test Items: Using Item Response Theory to Validate Assessments

Leigh Ann Sudol
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA
lsudol@andrew.cmu.edu

Cassandra Studer
Statistics Department
Carnegie Mellon University
Pittsburgh, PA
cstuder@cmu.edu

ABSTRACT
As professional educators we produce a large number of assessments for our students to complete. These assessments or exams are often evaluated informally based upon student feedback and simple measures such as average score on a question. This paper highlights another more rigorous approach to item evaluation, and presents an evaluation of several items from an assessment as examples of the type of information that Item Response Theory can provide.

Categories and Subject Descriptors
L.0.0 [Science and Technology of Learning]: Assessment/Evaluation/Measurement—Assessment and Evaluation; K.3 [Computers & Education]: Computer and Information Sciences Education; I.6 [Computing Methodologies]: Simulation and Modeling

General Terms
Item Response Theory

Keywords
Item Response Theory, Assessment

1. INTRODUCTION
There is an increased call for data-driven instruction within educational circles. In the speech that Secretary Arne Duncan gave to the Institute for Education Sciences this past June he called for an increase in research and use of data to direct classroom practices.[1] While there are a number of efforts at the national, regional, and institution level to use data as a tool, even individual teachers and professors can think about using data from within their own classes to improve upon instruction and in doing so improve student learning.

While assessments give us insight into student knowledge and outcomes, they can also be used to help inform us of the alignment between our goals and the instruction provided. On an assessment where a large percentage of students get a particular question incorrect, there could be many reasons why this occurred. The students just did not understand the material, the question was misunderstood, the question was too hard for the population, or perhaps relied upon some other underlying knowledge that wasn’t apparent. Statistics, and in particular Item Response Theory can be used to help us determine which of these situations is happening.[2]

In the past 10 years Item Response Theory has been used to validate assessments as well as help researchers understand how question difficulty and student performance interact in order to design better assessments.[3] Within education we are seeing an increase in the use of research designed tools in everyday classrooms. Open source content management systems like Moodle are becoming increasingly more popular in classrooms. As of September 2009 there were almost 38,000 registered sites using the Moodle platform alone.[4] Open source allows for the addition of tools such as item analysis to become available to the educational community. Web environments for data analysis such as the PSLC Data Shop can also be used for the tools they provide.[5]

Previous contributions to the SIGCSE community have advocated the use of Item Response Theory[6] for program level assessment as well as item validation within an instructional tool.[2, 7] This paper offers a case study in the use of Item Response theory to analyze a mid semester assessment with the purpose of improving the assessment for future semesters. In addition to example code from the R statistical package, exemplar graphs and their analysis are included within this paper. The results are shown not to highlight the particular assessment used, but instead to illuminate the process and offer guidance for others who wish to validate their own assessments.

2. WHAT IS IRT?
A student’s correct or incorrect response to an exam item is typically a mixture of his proficiency in the area that the exam is covering and the difficulty of the particular item. Item response theory[8] is a method that attempts to enumerate these student and item characteristics in order to calculate the probability that the examinees will answer correctly. Two currently used IRT models are the 1- and 2-parameter logistic (1PL & 2PL) models. Note that we will assume all questions are graded right/wrong but that the methods can be adapted to account for partial credit scoring without difficulty.
The simplest IRT model is the 1PL or Rasch model. The probability that student $i$ correctly answers item $j$ is modeled as,

$$P_j(\theta_i) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)},$$

where $\theta_i$ is student $i$'s proficiency and $b_j$ is the difficulty of item $j$. Higher values of $\theta$ indicate higher proficiency and higher values of $b$ indicate more difficult items. In order for a student to have a probability of 0.5 of answering an item correctly, he must have a proficiency equal to the item difficulty, i.e. $\theta_i = b_j$. Student proficiencies are often assumed to have a normal distribution with mean 0 and variance 1. Therefore, values typically fall between $-3$ and 3. Values outside of this range may indicate a problem with the item.

In addition to problem difficulty, we might want to include information about how well a particular problem differentiates students of varying abilities. We can add a discrimination parameter, $a_j$, to the model. This model, known as the 2PL, is modeled as,

$$P_j(\theta_i) = \frac{\exp(a_j(\theta_i - b_j))}{1 + \exp(a_j(\theta_i - b_j))},$$

where $\theta_i$ and $b_j$ are the same as the 1PL and $a_j$ is the discrimination of item $j$. Values of $a_j$ are typically between 0 and 2 with higher values indicating items that better differentiate students. One can note that if $a_j = 1$ for all $j$, we have the 1PL model. Similar to extreme difficulty parameters, items near the extremes and outside the range of 0 to 2 may indicate problems with the content or wording of the item. Specifically, negative discrimination parameters indicate that students with high proficiency are actually less likely than students with lower proficiency levels to answer correctly. Values near 0 mean that the item may depend on a construct other than the one intended. Discrimination parameters near 2 and above can indicate that an item depends on a “you know it or you don’t” trick or fact. These items should be reviewed to make sure they are typical of the type of tasks students should be able to perform.

### 2.1 Item Characteristic Curves

We can summarize an item’s properties by looking at its item characteristic curve (ICC). ICs represent the probability of correctly answering an item for different values of $\theta$ and are logistic curves bounded by 0 and 1. The difficulty of the item can be visualized as the point of inflection on the proficiency ($\theta$) scale. Curves shifted to the left (right) indicate easier (more difficult) items. Alternatively, the discrimination parameter is depicted by the slope of the curve. For reasonable values of $a_j$, steeper slopes indicate items that are better able to differentiate between students. Negative and flat slopes correspond to negative and low values of $a_j$, respectively. ICs that approach a step function indicate particularly high values of $a_j$.

### 2.2 Warnings

While IRT is a great method for analyzing the reliability and validity of exam items, it should not be used as a method of assessing students. The estimated proficiency parameters will align identically with the percent correct on the exam. Therefore, a poorly written exam question will effect both a student’s grade on the exam and his $\theta$ value as well. Also, the proficiency estimate is exactly that, an estimate with a standard deviation. Therefore, one should not be overly dependent on the values. There is little advantage to using $\theta$ values in place of exam scores to assess students.

The other warning that we offer is to be aware of the number of students taking the test versus the number of parameters being measured for each item. Trying to estimate too many parameters with too few data points can lead to serious problems of which software designed to perform IRT may not warn. While there is no absolute answer to how many students are needed to ensure valid estimates, there is a general rule of thumb. For teachers who are simply trying to validate their own exams, one can assume that five students are needed for each estimated parameter in the model. This means that ten students are needed to measure the mean and variance of the proficiency distribution and an additional $5 \cdot x \cdot J$ students are needed to estimate the item parameters in the $x$PL model where we assume the exam has $J$ items. However, more data points will always lead to more precise estimates of the parameters.

### 3. DATA AND CODE

This example analysis uses a mid semester exam from an introductory computer science course. The main focus of the exam was to assess the students’ knowledge of inheritance and interfaces. Although the exam had several parts, only the first part with short answer questions is used for this paper. There are graded response models available, however this paper is focused on the simpler case where we assume questions are marked as either correct or incorrect.

In order to prepare the data for use with the IRT models, it was entered into a spreadsheet. Within the spreadsheet each row corresponds to one student’s answers. Each cell is labeled with either a 1(correct) or 0(incorrect) for the student’s response to that question. Cell A1 therefore contains the first student’s answer to the first question. There are no row or column headers. The file was then saved in .csv format to be imported into the R statistical program.

#### 3.1 R Code

The following are code snippets that can be used in order to import your data and run the IRT models in the statistical package R. Please note that the # sign denotes a comment.

```r
library(ltm)

# Load the library used for IRT
examdata <- read.table(file.choose(), header=FALSE, sep=";")

# Run the model and store in result variable
result <- ltm(examdata ~ z1)

# Display the results (general text output)
summary(result)
```

```437```
#display the ICCs, the first line sets up
#the window for how many graphs (5 x 3)
par(mfrow = c(5, 3))

#print the graphs
for(i in 1:15)
plot(result, items=i, legend = TRUE,
     cx = "bottomright", lwd = 3, cex.main = 1.5,
     cex.lab = 1.3, cex = 1.1)

The above code for R is merely an example to help you get started. There are various references for R and the ltm package available online.

4. RESULTS

The following section discusses the item characteristic curves that were produced by the two parameter logistic model. This particular exam was chosen for this paper because it produced examples of many of the types of curves that can be encountered. When looking at these graphs the x axis represents the proficiency of the student, \( \theta \), and the y axis represents the probability that a student of the given proficiency is able to answer the question correctly. Using figure 1, a student with a mean ability for the class has a 60% chance of getting the item correct.

4.1 Target Curve

Figure 1 represents the desired shape of a valid assessment item. The steep slope indicates that the item does a good job of discriminating between students.

---

4.2 Guttman Items

Figure 2 also represents a desired shape, and is sometimes referred to as a Guttman Item since it is clear that there is some knowledge construct that is specifically responsible for answering the question correctly. Guttman items are often observed in exams of knowledge recall such as vocabulary tests.

---

4.3 Easy Questions

Sometimes items are created to be easy and will not discriminate among students within the class. These items are generally represented by curves that exist in the top half of the graph. Figure three contains some easy item examples.

---
4.4 Linear Items

Some questions produce Item Characteristic Curves that are linear in shape. These curves indicate that while students with more knowledge are more likely to answer the question correctly, there is something else going on. If the question relied on one, or a few specific skills then you would see the "S" shaped curves from figure 1. These questions should be re-evaluated for what knowledge is required to answer the question correctly.

The question that corresponds to this graph is:

```
Given the API for the Timestamp and Date objects, is the following a valid declaration?

Cloneable c = new Cloneable(...);
```

In this case we can hypothesize that it is not only a knowledge of inheritance here that is prompting the correct answer, but also the knowledge that you cannot construct an instance of an interface. The combination of those two pieces of knowledge are required to answer the question correctly and so the question is not good at discriminating among students with regards to their knowledge of inheritance, which is what most of the remainder of the exam tests.

4.5 Descending Curves

A descending curve such as figure 5 is a strong warning sign to anyone engaged in this kind of analysis. The descending curve indicates that students who had an better score overall had a smaller probability of answering the item correctly. If this pattern is witnessed then a careful evaluation of the question should be performed in order to determine the cause.

The question that prompted this curve was:

```
Consider the following constructor for the Timestamp class:

public Timestamp(long time){
    //code here
}
```

In order to answer this question students needed several concepts. First they needed to identify that the call to super was on the first line, determine that it would call the constructor for the Date class, and then look at the reference for Date (provided) to determine if there was a constructor which accepted the value 2000. There are many reasons why this ICC could be descending including requiring multiple pieces of knowledge to complete the question correctly. As I rewrite this exam for future use, I might ask the question in a different way, or in multiple parts so that it told me exactly what the student misunderstood.

5. GRADED RESPONSE MODEL

The models presented here work with assessment questions that are binary in scoring - i.e. correct or incorrect. Additional models are available if the questions are graded with possible partial credit. We anticipate a future paper detailing the graded response models as well as evaluating the results as a data set becomes available.

6. SUMMARY

This paper has presented you with an introduction to Item Response Theory and its potential uses for classroom level assessments. The included code as well as graph analysis is meant to provide the reader with a foundation to build upon as they become more engaged in the practice of assessment evaluation.

There are many other potentially useful tools that assessment validation research can offer such as Item and Test information curves, graded response models, and Knowledge Component models. There are also a number of other possible software tools aside from R that can be used to generate these curves as well. This has been a gentle introduction to a small, but potentially useful part of Item Response Theory that hopefully can be applied in the classroom quickly.

There are two important concepts to keep in mind as educators make use of these tools. First, the Item Characteristic Curves tell you more information about the questions than
the students. The x axis is dependent totally upon the student’s performance on that assessment and is normalized only for the data it has on the current students. Secondly, while the different shape curves indicate student response patterns, it is up to the instructor to evaluate the questions that generated that pattern and decide what steps to take next.

As data becomes an increasingly larger part of the arsenal that educators wield, so too must we become more sophisticated in the analysis and use of that data. It is the authors’ hope that this paper offers a gentle introduction into a powerful topic for the average practitioner.

7. ACKNOWLEDGMENTS

The authors would like to thank Ken Koedinger and Brian Junker for running an item response theory workshop which prompted the idea for this paper. This work was supported through the Program for Interdisciplinary Education Research (PIER) at Carnegie Mellon University. PIER is supported by a grant from the Institute for Education Sciences in the US Department of Education.

8. REFERENCES