EARTH & OCEAN SCIENCES SCIENCE EDUCATION INITIATIVE
The University Of British Columbia

The EOS Questions Database Project, Phase I Report, Eva Shaffer, August 2008.

Contact: Francis Jones, http://www.eos.ubc.ca/public/people/faculty/F.Jones.html

Contents

Introduction

Hello, and welcome to the Earth and Ocean Sciences Questions Database. This document outlines the database structure, use of the database, problems and challenges, possible uses for the database, an introduction to Item Response Theory (IRT) and the challenges associated with it, implications of IRT, and possible directions for research using IRT.

The database was created with multiple purposes in mind:

- To provide a place for storing and tracking multiple choice items used in introductory earth and ocean science classes
- To allow instructors to search for pre-existing items by module, concept, and difficulty
- To provide a tool for facilitating research into the overall effectiveness of introductory courses

- To further understanding and research into Item Response Theory (IRT) and how it might be used to analyze course effectiveness
- To aid instructors in creating more effective multiple choice tests
- To allow instructors to compare class performances on items that have been used multiple times

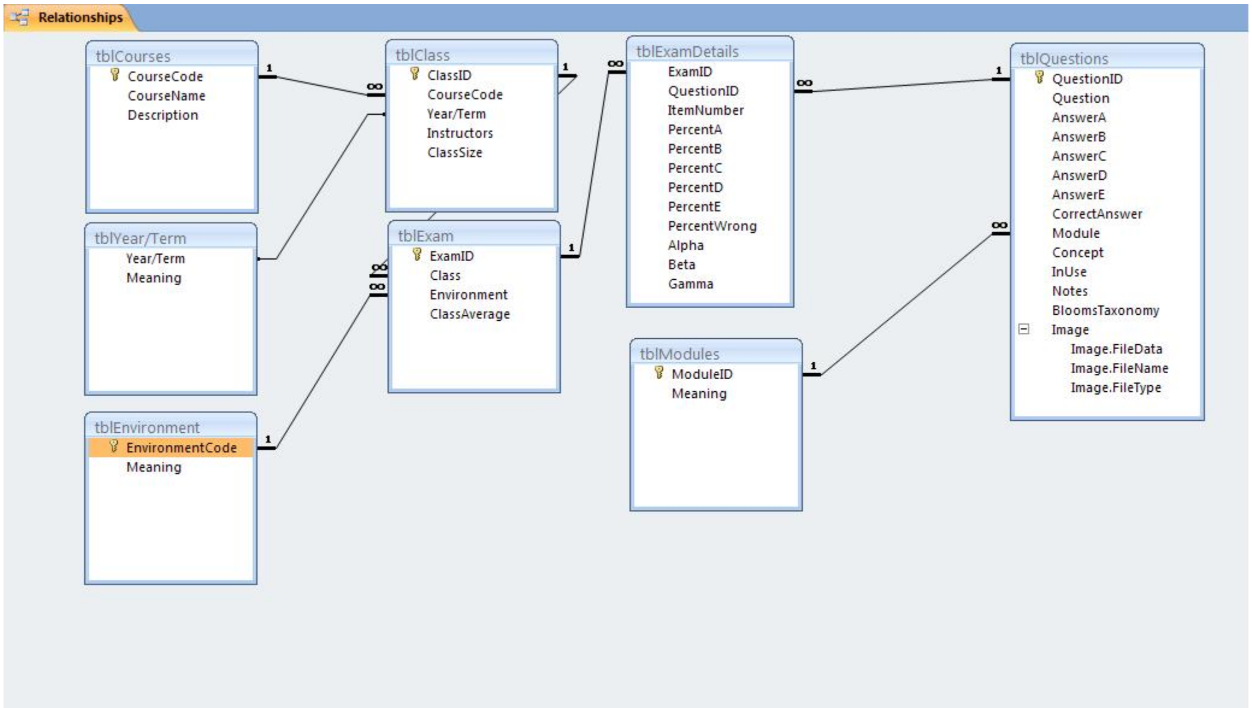
The items within the database will be taken from finals, midterms, and clicker questions administered within the introductory earth and ocean science courses. Instructors will be able to search for questions with a variety of parameters, as well as add new questions they have created. Data concerning class performance and distracter effectiveness (reflected by the percentage of students choosing each distracter) will also be added for every use of the question. This will allow instructors to compare performances between classes and exam settings. Preliminary IRT data will be connected to each question, but there are some limitations to these data that will be outlined in this guide.

At the completion of this first phase of the project, the database consists of several tables containing items and their associated data from the course Natural Disasters (EOSC 114) over the span of fall 2005 up to the present. This does not include items from summer school or distance education courses. These tables exist as a relational database within Microsoft Access. Details of the database structure and design will be described in the Layout section, and will serve as the foundational design for the implementation phase.

As a second phase of the project, the database will be implemented with help from EOS computing staff on the departmental server with access being provided via a web-base interface. The intention is to use a pre-existing my.eos account name and password to log in and use the database. Each instructor will be granted a level of access dependent upon his or her role in the course. For example, all questions will be open to search queries, but only instructors with the authority to set exams will be able to add new questions or edit pre-existing questions.

The Database Structure

The structure of the database is illustrated in img1, which shows the relationships between separate tables within the database, as it was designed in Microsoft Access.

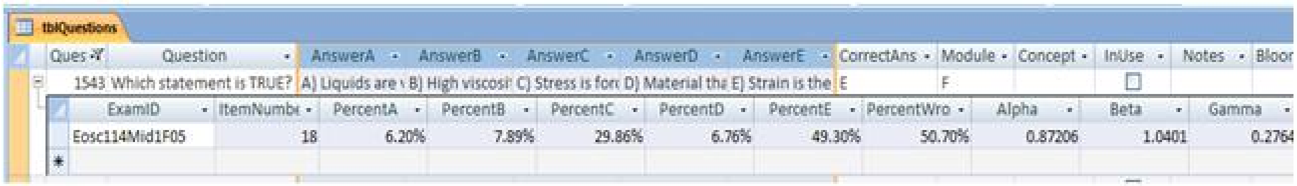


IMG1: Relationships in the EOS QDB. As designed in Microsoft Access.

Questions Table

A table exists within the database that documents all the unchangeable aspects of an exam item. This data includes the question itself, each distracter, and the module it belongs to. There is an area for any stray notes that an instructor may wish to add to the question, and images can be attached to each item as well. This is especially important, as introductory EOS exams are becoming more image based. There is also a toggle field to indicate whether or not an instructor is currently using the question for an exam. This will prevent two courses using the same item at the same time, which is important if students are taking more than one introductory course.

As well, there are fields for “Concept” and “Bloom’s Taxonomy”. This is envisioned as something instructors will edit as they add more items. Pre-existing items have not been ranked using Bloom’s or separated by concept. Instructors with a better knowledge of their course content and how the items have been structured would be better able to fill in these fields. Each item is given a unique question ID number. This process is automated; a question is assigned an ID number when it is added to the database. Img2 Shows a sample question with its related fields. As well, an expandable menu shows information that can be found in the Exam Details table.



Ques	Question	AnswerA	AnswerB	AnswerC	AnswerD	AnswerE	CorrectAns	Module	Concept	InUse	Notes	Bloor
1543	Which statement is TRUE?	A) Liquids are \	B) High viscosi	C) Stress is for	D) Material tha	E) Strain is the	E	F		□		

ExamID	ItemNumbe	PercentA	PercentB	PercentC	PercentD	PercentE	PercentWro	Alpha	Beta	Gamma
Eosc114Mid1F05	18	6.20%	7.89%	29.86%	6.76%	49.30%	50.70%	0.87206	1.0401	0.2764
*										

IMG2: sample question in question table showing all fields, as well as displaying information from the related exam details table.

Exam Details Table

Data about class performance and IRT can be stored for multiple uses of a single question in a separate table. This separate table is the “Exam Details” table. Within this table, the class’s performance and distracter effectiveness – as signified by percent wrong and percent of students to indicate each distracter as the correct answer – are recorded for each use of the question. The item response theory parameters (alpha, beta, and gamma) that have been calculated for each question are also recorded in this table. As mentioned above, these parameters will be discussed in a later section. Through this design, it becomes possible to compare different uses of the same question, such as how the question performed in a midterm versus how it performed in a final, or whether or not one course covered the material more effectively than another course.

Exam Table

Each separate time an item is used, it is used within a different setting. This could be as a midterm, final, or a clicker question, in a different year, or for a different course. This discrepancy is accounted for by including with every separate use of an item an Exam ID code. This code is generated from information about the class and what type of assignment it was administered as, all of which exists in a separate table. Within this table, information on the class average for the exam is also stored. The Exam ID’s are unique for each separate exam.

Class Table

Classes are delineated with their own Class ID, which is generated from information about the course and the year and term of the class. This data is also stored in a separate table, which also includes data on the class size and instructors for that class.

Satellite Tables

There are several satellite tables that store information regarding meanings for shorthand. These allow for referential integrity, and the creation of dropdown lists. The database is better capable of consistent data entry with fewer errors when separate tables exist to keep track of what short-hand for modules or course codes are.

Ideas and Concerns

The final details of what fields will exist in all tables will only be finalized after the database has been through several design-test-modify cycles. Currently there are several issues and suggestions still to be considered:

- Addition of text fields for feedback given to students after they complete the question. This is useful for online deployment of questions sets, and could include a single feedback message for the question and/or a feedback message associated with each distractor.
- We have not yet decided how to handle the possibility that instructors may want to include questions with more than 5 options. This is known to be done for some clicker questions used in EOSC 114.

- There is currently no mechanism other than the notes field to identify questions that are coupled. For example, some clicker question activities may include more than one question used back to back.
- Instructors need to be encouraged to include notes on pedagogy and usage of questions if they are intended for active learning situations rather than testing or exams. This is one application for the notes field, but instructors need to be encouraged, perhaps with a suggestive message provided as part of the data entry screen.
- There is currently no "title" field for each question (as is commonly used in other systems), but see a note on this at the end of the "Problems and Challenges" section.

Using the Database

There were many uses for the database listed at the beginning of this manual. In its current form, users who are familiar with the MS Access Database program can carry out all data entry, querying, and report production using the built in facilities of the MS Access system. Custom forms for carrying out these tasks have not been completed owing to a shortage of time and the fact that the final version is intended to be implemented as an online facility so that all instructors in the department can make use of the database.

The next four sections are an attempt to walk you through the forms and queries associated with each use, as well as simple start up of the database.

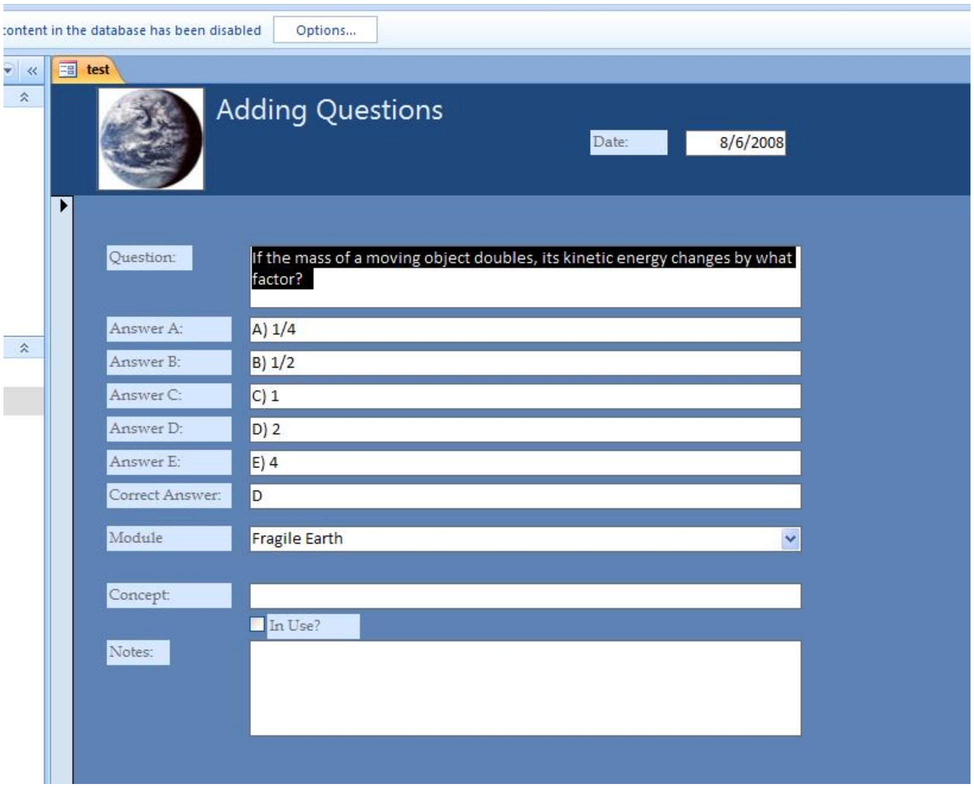
Start Up

As an object on the my.eos server, we hope that instructors will either need to log in to their my.eos account, or access the page that is hosting the database and use their my.eos login name and password. Upon logging in, you should be presented with a "switchboard" that will present you with buttons representing options that are available to each user. There will be an "Add Item" option, a "Build Test" option, and a "Search" option.

Adding Items

Clicking the "Add Item" option will bring you to a new window. Within this window you will be able to choose from a menu what you would like to add: a new question, a new course, a new exam, a new module, or new exam details. It should be noted that there would likely be limitations on each account. For example, only someone with an administrator status would be able to add in new courses, and only instructors for courses will be able to add exams or classes for that course.

Upon choosing what table you would like to add an item to, you will be brought to a new window. This will contain a form with spaces for all of the information that you will add to the table. For example, if you were to click on the option that allows you to add a new question, you would be presented with a form (Img3) that would have fields for the question, distracters A through E, the correct answer, the module and concept the question belongs to, whether or not the question is in use, notes, and an attachment button so that you could upload an image if you needed. Simply fill out these feels and click the "add" button. Similar forms exist for adding a new course, a new exam, and a new module.



Date: 8/6/2008

Question: If the mass of a moving object doubles, its kinetic energy changes by what factor?

Answer A: A) 1/4

Answer B: B) 1/2

Answer C: C) 1

Answer D: D) 2

Answer E: E) 4

Correct Answer: D

Module: Fragile Earth

Concept:

□ In Use?

Notes:

Img3: Adding Questions to the EOS QDB. As designed in Microsoft Access.

The only anomaly within the add function is with the exam details. As upwards of fifty items will need to be added in a single go, we are working on an interface where a properly formatted text document will be able to be uploaded and the items added en masse. It is likely that this properly formatted document will simply be the Scantron output for the questions in the exam.

One problem with this is that there is a tendency to randomize questions between different versions of the exam. This will mean that all versions of the exam will need to be collated, and then matched up to the original question ID number they were given in the question table. This will be discussed further in the problems section.

Building a Test

Clicking the “Build Test” button will bring you to a new window. This window will ask you some specifics, like which course you would like to create an exam for, and which modules you would like the exam to cover. This will limit the following search criteria to a specific set of items relating to those parameters.

Then you will be able to specify which concept for the course you would like to search for. You will also be able to add in specific words or phrases within another field, if, for example, you want something on “Volcanoes” but you would also like the question to be about “British Columbia”. If you wish, you can also specify the difficulty level as defined by item response theory, classical theory[1], or Bloom's taxonomy levels (Blooms, 1956). For IRT and classical theory, you will be able to choose from a ranked scale of very easy, easy, moderate, hard, and very hard. For classical theory, each level of this scale will be connected to an associated range of percentage wrong. For example, "very easy" would be linked to questions where 0-20% of the class answered incorrectly; "easy" would be linked to questions where 21-40% answered incorrectly, and so on. For IRT, the scale would operate in a similar way, but using the difficulty parameter "alpha" as its reference, rather than the percentage that answered incorrectly. The Bloom's taxonomy scale would simply use the six levels already set out within Bloom's: knowledge, comprehension, application, analysis, synthesis, and evaluation.

A problem exists here in that different classes will perform differently under classical theory depending upon the original testing environment. As well, this problem exists within IRT. This will be explored further in the problems section. For obvious reasons, searching for items using Bloom's Taxonomy as a parameter will not include questions that do not have data within that field.

Once the parameters for your search have been defined and you have clicked the "Run Search" button, all items fitting within those parameters will come up in a table. You can read through these items. There will be a toggle box next to each item. You can toggle this box on or off depending upon whether you want to use the question or not. This is very similar to how a shopping cart works in an online store. Once you have indicated all the questions you would like to use from this search, you can indicate whether you would like to finish the exam or continue with another search. In this way you can run searches for several difficulty levels, modules, or concepts.

Once you have run your searches and indicated all of the questions you would like to use, you can click the "Next" button at the top right of the page. This will bring you to a new window that has all of the items you chose listed next to a field of toggle boxes. Here you can make your final decisions about which questions to use, using the toggle boxes to indicate questions you would like to remove from the exam and then clicking the "remove question" button.

Once you are happy with your selections, you may click the button indicating "Finish Assignment" button at the top of the page. This will cause the database to create a text document. The document will contain only the questions and their distracters. This document will be formatted so that it is ready to upload into Respondus[2]. If you wish to administer a paper exam, it is as easy as reformatting the text document yourself, or with the help of Respondus.

[1] Classical test theory is based on the decomposition of test outcomes into true and error scores. Classical test theory does not rely upon the fixed property of tests, but on the on the properties of tests scores that are dependent on a particular population. For the database, percentage incorrect scores were used to indicate the difficulty of the exam questions relative to the population of the students taking the exam. It should be noted that classical test theory becomes more accurate as population sizes increase. For more information, please visit: http://en.wikipedia.org/wiki/Classical_test_theory

[2] Respondus is a program available to all UBC faculty which will communicate with Vista so that you may upload or download your exam onto the internet for students to take. It is in fact a sophisticated question management system in its own right. See the UBC eLearning website at https://www.elearning.ubc.ca/home/index.cfm?p=main/dsp_respondus_index.inc for details.

Searching

Searching for items would be used for finding or comparing individual questions, or for doing research on how individual (or sets of) questions are serving the learning needs of students. It will work in a very similar way to the "Create" function, but will not end in outputting a report of the questions you have found. This could simply be a way to find a question and compare its uses, or you could search for questions with similar terms and compare their difficulty levels.

Once you click "Search", you will be brought to a new window with a variety of fields to fill out. Not all of these fields are required. There will also be a field of toggle switches where you can indicate which data you would like to see accompanying your search. If you know the question ID of the item you would like to bring up, you can type this in. It will bring up that question, and listed below will be all the data that is attached to that question that you have indicated that you would like to see, such as distracters, correct answer, and module, as well as the performance data for any exam that used it, including any item response theory data that has been added. In this way, it will be easy to compare multiple uses of a single item.

If you are looking for a specific module, or specific terms, you can type those terms into the "By Keywords" field. This will bring up all items that contain those keywords. These questions may be very similar, or quite different. All the data that you have indicated in the toggle field prior to running the search will also be revealed. In this way, it will be possible to analyze how variations of similar questions may have an effect on how well students perform, and how that affects the item response theory parameters.

Searching through modules or by difficulty level for specific classes can yield data on how students are performing on each separate module. You can even bring up an entire exam and analyze how well students performed on different aspects of it, such as by module or concept. It will even be possible to bring up all questions used in the same term and year. Any combination or variation of data that exists within the database will be fully searchable.

Problems and Challenges

One of the problems that became apparent as the project progressed was a lack of data on same and similar questions. Questions within EOSC 114 (the test course for this project) are often re-used in different testing environments, after being tweaked slightly by the instructor. This is often an attempt to change the question from a past use, or to perfect the question, or to put a spin on the question that is more relevant to the class. How this fits into the database is as yet undecided. Duplicate questions have been left in the database, but it is hoped that for future uses, instructors will keep track of which questions they use, and make sure that test response is directly uploaded to those questions, rather than re-entered as a new item. It is hoped that instructors will record which items are derivations of other items in the "notes" field of the database. This way it will be easier to keep track of how items have evolved, and what affect that has had on a class' performance.

The most useful aspect of having duplicate questions within the database was to indicate that the IRT data was not the same for the same question administered in different environments. This emphasized that there were problems inherent in using IRT, and this will be outlined below. It also led to the decision that IRT data should not be stored with the "unchangeable" aspects of an item, such as the question wording, distracters, and module it belonged to. Rather, it was stored with the data that is considered “changeable” depending upon the class and environment, such as class performance and distracter effectiveness.

Another problem is that it is necessary to create a template for uploading a large amount of data at one time. There is a tendency for instructors to create one exam, and then randomize the questions for different versions of the same exam. This means that when uploading the data, it becomes more difficult to make sure that the proper information is being assigned to each question. It becomes necessary to introduce more steps into the uploading phase. A possible solution to this is to ensure that within each version of the test, a record is kept of the question’s unique ID number. That way, each version of the test can be uploaded as a separate test with a separate test ID number, and each question item will still be linked with the appropriate data. One way of implementing this would be to include the question ID number as the question’s “title” during export. This title does not need to be included in printed reports (i.e. tests), but both Respondus and WebCT/Vista can accommodate an optional title as part of each question.

Possible Directions

There is difficulty in creating one database to fulfill multiple purposes for multiple people. Ideally, as the project evolves, some level of personalization for the database will emerge. Instructors will be able to save specific queries, or alter different aspects of the database to suit their needs. These could then be saved onto their account without affecting the usefulness of the database to others.

There is the possibility of taking the data from same and similar questions and comparing how the different iterations of each question performed. It is difficult to fully understand how minor changes to a question can alter the questions’ ability to assess students effectively, considering the large amount of variables that exist between classes, courses, and environments, but having all the data in the same place helps. The lack of controls over how they were initially administered lessens the usefulness of existing data, but it points to how we could gain more useful data.

Suggestions include limiting the number of items in an exam in order to prevent speededness from affecting the class outcomes, as well as utilizing clickers or classical test theory data from past tests to pinpoint common misconceptions and then design items around those misconceptions.

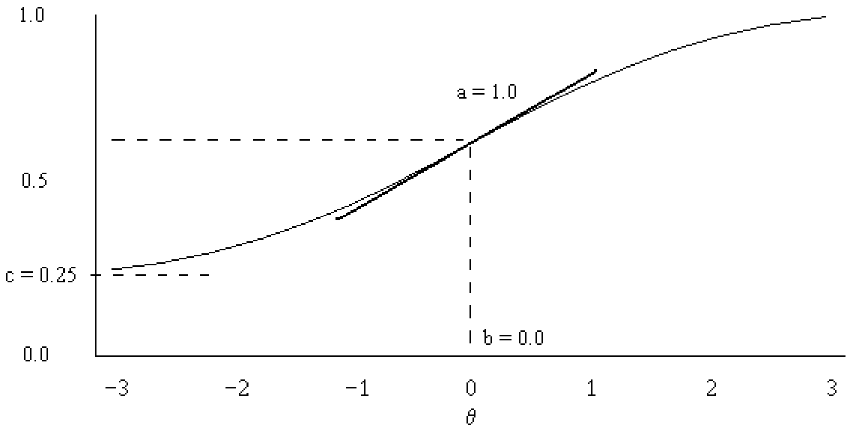
Coming up with a means of tracking a question’s evolution and its resulting effectiveness over the years would be an interesting way to utilize this database. This way, the faculty can work towards a set of questions that can be held as an example of highly effectual, well created test questions that have a useful body of data surrounding and supporting their ability to effectively assess students’ abilities. With a model of a good test question and knowledge of how it was arrived at, as well as knowing what worked and what didn’t, creating new items may become easier and more streamlined. Keeping track of a question automatically raises an instructor’s awareness of how that question has performed in the past. As well, the knowledge base will be there for analyzing a question’s effectiveness.

Evolution is the key for a project of this size. Starting off with something small and malleable and then working up to something robust and ultimately useful is the goal of this project. It is hoped that output methods, item search and assignment creation tools, and the user interface will become easier and more intuitive to use. Creating something like this is an iterative process that needs plenty of feedback from its users. To maximize the usefulness of this project, inputs in the form of suggestions and thoughts once a beta version is up is encouraged. This way, different aspects of the project can be fine-tuned and tweaked to perfection.

Item Response Theory

Item response theory (IRT) is a powerful assessment tool that more and more people are using to analyze exam structure. Unlike classical theory, it provides parameters for questions independent of the ability level of the examinee. It is commonly used for unidimensional psychometric testing and large-scale, high-stakes assessments such as SATs, but is not usually applied to university assessments.

In IRT, analysis of exam data will provide each question with an “item characteristic curve”. This represents the item’s difficulty, which describes where the item functions along the ability scale, and the items discrimination, which describes how well an item can differentiate between examinees having abilities below the item location and those having abilities above the item location (Img4).



IMG4: ITEM CHARACTERISTIC CURVE

In the curve above, the Y-axis measures the probability that a question will be answered correctly and the X-axis measures ability level, and is often referred to as theta. Ostensibly, it goes from negative infinity to positive infinity, but for general use it tends to be limited, usually between +/- 3 or +/- 7, depending upon the software used to run the calculations. A student’s ability level is plotted along the X-axis (measured using theta). The curve therefore measures the probability that a student with a given ability level will answer the question correctly.

A program called PARDUX (CTB/McGraw-Hill, 1991) was used to estimate item parameters and characteristics[3].

IRT can be used to create “designer tests”, where an instructor can choose an ability level they wish to test at, and then choose questions that will test for that ability level, and discriminate as sharply as

[3] Many thanks to Jackie Stewart, a research associate (Joint with The Science Centre for Learning and Teaching (Skylight)) currently conducting research on Item Response Theory within the Chemistry Department at UBC. Her research focuses on use for organic chemistry (Chem 233) and large scale testing within the science department. Her aide in understanding IRT and using PARDUX was invaluable.

they wish between students above and below that ability level. For example, a medical school exam, where a high level of performance is required, would choose many questions at the upper end of the difficulty spectrum and few questions from the moderate or low end. These questions would preferable have a steep alpha parameter. In this way, only students who perform at a high theta level would pass, and students with abilities below the level being tested for would fail.

In the database, the item's ability to discriminate between students' ability levels is described by the alpha parameter. This measure is the maximum slope of the item characteristic curve. As may be apparent with the image above, a steeper slope, or a higher alpha parameter, means that the item has a greater ability to differentiate between students before and after the point at which it occurs. The point along the X-axis at which this maximum slope occurs is called the beta parameter, and indicates the item's difficulty level. The gamma parameter is the point at which an asymptote towards negative infinity occurs. This measures the probability that a student will guess the correct answer regardless of ability. In a multiple-choice environment, there is always the possibility that a student will be able to guess the answer.

For more information on the math underlying IRT two acceptable online resources are: http://en.wikipedia.org/wiki/Item_response_theory, and http://luna.cas.usf.edu/~mbrannic/files/pmet/irt.htm .

Challenges with Using IRT

As mentioned above, there are some problems with using IRT on pre-existing test data. These problems will be explored in depth below.

Coverage of Course Material

A problem with IRT is that if an entire class has been made less familiar with one concept over another, it may skew results. For example, consider an "easy" item concerned with an area of the course that has not been sufficiently covered at the time of the exam. The item may be given a high beta parameter (associated with a difficult question) when it should have a low beta parameter (an easy question) if a majority of the class answers incorrectly, while at the same time the students who answer correctly are also students who are performing well in other areas of the test. They may perform well because they are more familiar with the subject matter, outside of the class. The question may appear to be harder than it is in this way. It would seem classical test theory is as accurate as item response theory in picking out questions from areas that are perhaps not covered sufficiently.

Unidimensionality

Unidimensionality is needed for an accurate measurement of ability level. The specific ability being tested for (in this case, presumably, geoscience knowledge) must not be confused with an ability to read English, comprehend the question, or recall information quickly. Reading comprehension can be described of as a "nuisance ability" (Ackerman, 1992), or a skill which an examinee uses to solve a particular item but which the examiner had not intended to test. These abilities are nearly impossible to avoid and need to be taken into account. Ackerman notes that "researches must examine the conditional distribution of the nuisance ability for each level of the valid ability."

Understandably it is hard to have an item be 100% unidimensional. In these earth and ocean science exams, some level of reading ability and comprehension must be assumed, as well as fluency in English. In order to be able to run a unidimensional test in IRT, we have to assume that examinees only vary significantly on the same composite of skills. We must assume a base level of reading comprehension and make sure that questions aren't too "tricky" with the wording.

Ackerman (1992) outlines three ways in which an exam item may be unidimensional:

An item may be sensitive to, or require the application of, several skills to produce a correct response. But, if examinees only vary significantly on one of the requisite skills or on the same composite of skills, the interaction can appropriately be modeled unidimensionally. The reverse scenario is also possible: test items may be sensitive to, or capable of measuring, only a single skill or the same composite of skills. Then, although examinees vary along several skill dimensions, the interaction will be unidimensional. The third situation is the degenerate case in which the test is only one item long. Considered by itself, one item is always unidimensional. It is probably true that a test composed of two or more items is never exactly unidimensional.

For our purposes, we can only consider the first two. Realistically, we are likely dealing only with the first method of determining unidimensionality, and we must then include within our composite of skills sufficient reading comprehension and fluency in English to understand the questions be asked. Then the skill set being tested is the knowledge gained during the earth and ocean science course.

It would make sense that students from different faculties would be bringing different skill sets into the test. To generalize, a science student may be more familiar with the material being presented, or more comfortable with equations; an arts student may have the ability to read through questions quickly with a high level of comprehension, or may be able to articulate a long answer more concisely and accurately.

Speededness

In a study by Oshima (1994), it was shown that speededness affects the estimation of item and ability parameters. It can slightly distort the estimation of ability. However, the author found that the correlation between the true and estimated ability parameters was fairly high, and concluded that if the purpose of an exam is to rank students, violating a non-speededness assumption may not be too serious a problem. The author recommends using a program that will take into account not-present item answers. It was suggested that items near the end of a speeded test are more likely to deviate from their true values. In EOSC 114, however, the questions are randomized in three or four separate tests. This randomization may reduce the effect of speededness on items near the end of an exam.

The research above indicates that speededness may be one of the factors that could create differences when we compare item response parameters from the same item administered in different testing environments. Significant deviations of the same item's parameters from each other may be an indication that some other factor, such as speededness, was affecting the parameters of the test.

Another interesting note is that it is possible to indicate an estimation of the percentage of questions that are non-speeded within a test. This often applies to the first questions on the test (i.e. the first twenty questions on a forty question exam could be completely non-speeded).

Construct Validity

Testing for bias may reveal “construct validity” issues. A test has construct validity if it is measuring only the ability that the examiner wishes to measure. If a test has items that measure abilities other than those being tested for, a bias might exist. If groups of interest (perhaps separated by gender, race, or education background) perform differently on these items based upon some presumed underlying skill, bias may be detected. While it may be difficult to pinpoint what exactly the bias is based upon groups (gender, race, background, etc.) the fact that bias exists may point to the fact that a question is inaccurately measuring the ability being tested. Conversely, if all items in an exam are measuring only the desired skill or skill set, group differences reflect impact, not bias. Ackerman (1992) formerly defines impact as “a between-group difference in test performance caused by a between-group difference on a valid skill (e.g., the difference between the proportions correct for two groups of interest on a valid item).

Implications For IRT

Strictly speaking, with the proper controls and a lack of bias or speededness in the questions, the IRT parameters should remain the same for an exam item regardless of the environment it is administered in, or who the test is being administered to. At best there should be a statistically insignificant difference between the parameters of an item administered to two different groups of people. However, a cursory analysis of the IRT parameters for identical questions administered in different environments to different groups of people revealed that this was not true.

This suggests that the items themselves are not unidimensional, that they contain biases, or that the exams are speeded. This is a good start-off point for further research into IRT. Questions need to be tailored in such a way that they do not contain bias and that they are unidimensional, and then administered in an environment where students have the time they need to read, understand, and answer the questions. This is what was done for the Geosciences Concept Inventory (GCI) developed by Drs. Julie Libarkin and Steve Anderson (2006).

For the GCI, grounded theory and scale development theory were used to generate the items initially. Extensive interviewing and questionnaires were used to create attractive distracters and reduce the ability of a student to “guess” the correct answer. Common misconceptions were used as wrong answers for these multiple-choice questions.

After the initial list of questions was developed, they were tested for validity and reliability in a variety of ways. Wordings were changed, distracters altered, and finally, item response theory tests were run to measure what ability level these questions tested for, and whether or not they contained bias. Some questions contained gender biases, and some contained racial bias. These were then left out of the inventory.

It is apparent that this is an exhaustive process, and time consuming. Constructing multiple-choice questions that are non-biased and effective at measuring class performance and ability is a time consuming process. Adding IRT to this process would likely only make it longer. Analyzing pre-existing data is difficult due to the lack of a controlled environment, and all the problems listed above.

However, that is not to say that the data is not useful. Knowing that an item in different environments is producing different parameters leads to the question of why. By seeing questions that are not performing as expected, we can perhaps see which areas of the course are not being taught or tested for effectively. Using the classical test theory data alongside the IRT data to pinpoint anomalies, and then researching what is different about those questions, could aide understanding into effective exam preparation.

Summary and Direction For IRT

The design and creation of this database seemed most effective in pointing out shortcomings in the exam environments. If IRT is in fact a route that instructors would like to take, it is advisable that each instructor create a small amount of questions that are administered in a controlled environment. This is where the clicker tests for concept testing could come in handy. It could provide the data needed for creating effective exam items, such as common misconceptions, as well as a non-speeded environment. Creating only a small amount of items also limits time constraints and allows instructors to garner useful feedback from students. Conducting short interviews and "talk-throughs" with volunteer students might also be helpful.

Changing the course to incorporate IRT should at best be a gradual process. As classes get larger and larger, fair and unbiased assessment becomes more difficult. It is an unfortunate trade-off. As our understanding and ability to come up with more significant IRT data increases, it is hoped that IRT parameters will be included within the "questions" table of the database, that is, the unchangeable aspects of the question, such as the wording of the question and its associated distracters and module. It will also lead to creating a more sophisticated search for question difficulty, where an instructor with sufficient IRT knowledge will be able to design an exam that specifically tests for the level of ability that they wish.

There is also the possibility that an accurate ability level designation would allow for "smart testing", where depending upon a student's response to a prior question, they would receive either an easier or harder question the next time around. In this way you could accurately assess their ability level by which level they were consistently able to answer correctly at and which level they were consistently unable to get past.

Conclusion

We believe that Phase I of this Questions Database project has shown significant potential, both for improving effectiveness and efficiency of generating assessments, and for facilitating action research about design, deployment, and effectiveness of assessment. Of course there is plenty of room for growth. It is hoped that this will become a useful tool for all instructors.

Sources

Ackerman, T. (1992). *A didactic explanation of item bias, item impact, and item validity from a multidimensional IRT perspective*. Journal of Educational Measurement, 29, 67–91.

Bloom B. S. (1956). *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain.* New York: David McKay Co Inc.

CTB/McGraw-Hill. (1991). [PARDUX] . Monterey, CA: CTB/McGraw-Hill.

Libarkin, J.C., and Anderson, S.W. (2006). *Science concept inventory development in higher education: A mixed-methods approach in the geosciences*. Journal of Research in Science Teaching.

Oshima, T.C. (1994). The *effect of speededness on parameter estimation in item response theory*. Journal of Education Measurement, 31-3: 200-219

Shealey, R. and Stout, W. (1993). *A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF*. Psychometrika, 58-2: 159-194