ABSTRACT

New accreditation requirements focus on education as a “continuous improvement process.” The most important part of such a process is that information gets fed back into the system to improve the quality of output. This requirement is often interpreted to mean a feedback loop that iterates on offerings of courses or entire academic years. This paper provides two smaller and more immediate feedback loops based on information theory. These techniques give instructors feedback on the quality of each question on a test or quiz, as well as a numeric score for the instrument of assessment itself. A tool implementing these can be used to help instructors learn which questions are difficult, as well as what types of questions are correlated with ability, and how to design a meaningful instrument of assessment. Comparing these measurements against the percentage of students responding correctly shows a clear benefit for these new techniques.

Categories and Subject Descriptors
K.3.2 [Computing Milieux]: Computer and Information Science Education—Accreditation; E.4 [Data]: Information Theory

Keywords
Feedback, Continuous Improvement, Mutual Information, Assessment

1. INTRODUCTION

In the 1990’s the idea of “Total Quality Management” or “Continuous Improvement Processes” swept throughout the manufacturing and engineering communities. This idea is conceptually simple: rather than blindly continuing with the manufacturing process as usual, set goals and take data through the process, and then alter the process to address discrepancies whenever the goals are not being met. From a system design standpoint, the core concept is to introduce feedback mechanisms where once there were none.

The educational process is much like a manufacturing process: we have a continuous stream of outputs (students), and we have goals regarding the quality of our outputs (what we would like them to know or be able to do). EC2000[1], the requirements under which many CS programs are now evaluated, requires that we no longer blindly proceed on gut-feelings regarding what worked and what did not, but instead move toward more quantitative measures of educational effectiveness.

In many cases, the feedback loops being introduced into EC2000 compliant programs are high level: the workings of each course are examined, commented on, and reported for the next instructor of the course in the hope that the same mistakes will not be repeated. If there is sufficient “buy-in” on the part of the instructors involved, this has the potential to effect real long-term change. Courses will become more similar from offering to offering, and better calibrated to the students coming into the course.

There exist other levels in which a feedback method is useful in teaching a course, including which quizzes or tests (henceforth referred to as instruments or instruments of assessment) were “best”, and which questions on those instruments were most information-rich. Basic information theory can be leveraged in the development of techniques for answering these questions. By getting a numeric score and computer-generated suggestions for which sorts of questions (in terms of difficulty) would make a given test more accurate, it is our hope that instructors can learn to construct more accurate instruments. By performing question-level analysis, instructors can learn which sorts of questions to ask, and more importantly may discover which topics were incompletely understood by the class.

The rest of this paper is divided up as follows. Section 2 presents the notation that we will use as well as basic equations and concepts from information theory and educational statistics. We then build upon these concepts to evaluate items in Section 3 and instruments in Section 4. The result of applying our methods to real student score data are presented in Section 5. A summary and conclusion can be found in Section 7.
2. BACKGROUND AND NOTATION

For an instrument administered to \( m \) students consisting of \( n \) questions, let \( Q \) be an \( m \times n \) boolean matrix where the entry \( q_{i,s} \) denotes whether student \( s \) got the question \( i \) right or wrong. Let \( q_i \) be the vector of scores for question \( i \), and \( g_s \) be the vector of scores for student \( s \) -- the usage will be obvious from context. Additionally, we define \( g_s \) to be the discrete course grade for student \( s \) given some discretization (pass / fail, A / B / C, etc), and \( g \) to be the vector of all such grades. Similarly, let \( \theta_s \) be the (real valued) score in the course (usually a course percentage) and \( \theta \) be the vector of all values of \( \theta_s \).

2.1 Information Theory

This paper will build on two major concepts from information theory: entropy and mutual information[3]. To properly present mutual information in our context we also need to introduce the concept of joint entropy. The entropy of a discrete random variable \( X \) whose possible outcomes are \( \{x_1, x_2, \ldots, x_m\} = \mathcal{X} \) with probability distribution \( \Pr_X \) is defined to be

\[
H(X) = -\sum_{x \in \mathcal{X}} \Pr_X(x) \log_2 \Pr_X(x).
\]

Given a set of samples for which \( \Pr_x \) is not known, the distribution can be approximated for each outcome \( x \) as the percentage of the samples that were observed as outcome \( x \). Similarly, we can estimate the entropy of the distribution with samples from that distribution:

\[
\hat{H}(X) = -\sum_{i=1}^{m} \frac{c(i)}{n} \log_2 \left( \frac{c(i)}{n} \right),
\]

where \( c(i) \) is the count of the observations of outcome \( i \) and \( n \) is the total number of observations. This is known as the empirical entropy.

Intuitively, one may view the entropy of a collection of observations as the minimum number of bits needed to encode one of the observations on average. Thus this measurement is the same as the average Huffman encoding length of a set of symbols, to use a more familiar example from CS pedagogy.

The entropy of two random variables \( X \) and \( Y \) is calculated in the same way, simply by taking the observations of \( X \) and \( Y \) jointly (the first observation of \( X \) with the first of \( Y \), second of \( X \) with the second of \( Y \), etc) and calculating in the same fashion. More formally this is

\[
H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \Pr_{XY}(x,y) \log_2 \Pr_{XY}(x,y),
\]

which extends in the obvious way to more than 2 variables.

The mutual information of two random variables is a measurement of how many bits are shared between a sample of \( X \) and a sample of \( Y \). It captures the correlation between two random variables, regardless of the mathematical relationship between the two. Thus it is not important to know how \( X \) and \( Y \) are correlated, \( I(X;Y) \) captures how strongly related they are. It is defined as

\[
I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \Pr_{XY}(x,y) \log_2 \frac{\Pr_{XY}(x,y)}{\Pr_X(x)\Pr_Y(y)} = H(X) + H(Y) - H(X,Y),
\]

where the second line follows naturally from the definition of entropy. If we are looking for the values \( X \) and \( Y \) to be related, a larger value for mutual information is better.

2.2 Item Response Theory

Within the realm of computer-aided assessment, one of the most powerful tools is item-response theory (IRT)[2], the power behind such texts as the computer-based GRE[4]. The basics of IRT can be summarized briefly. Assume the object of assessment is to estimate on some arbitrary quantitative scale the level of aptitude \( \theta_s \) for each student \( s \) in some area. Obviously \( \theta_s \) is not something that can be directly measured: no meter stick, balance, or mass spectrometer can tell us exactly how much a student knows about linked-lists. Therefore \( \theta_s \) must be measured indirectly for each student, and is thus known as a latent trait. We will use “latent trait” and “ability” interchangeably. Each question is assumed to have some “characteristic curve,” a plot of \( \theta \) versus the probability of a student with \( \theta = \theta \) answering the question correctly. There are several equations that are used for characteristic curves, but we are here only concerned with the concepts from IRT. Most of the common equations incorporate two parameters: \( \alpha \), the discrimination factor, and \( \beta \) the difficulty. Intuitively, \( \beta \) governs where the 50% level for a question lies along the ability/difficulty axis, and \( \alpha \) governs how steep the curve is at that point.

A question with a perfect \( \alpha \) will look like a step function and have the property that every student with an ability level greater than \( \beta \) will get the question correct, while every student with an ability level less than \( \beta \) will get the question incorrect\(^1\). Thus a student’s performance on such a question indicates without doubt how their ability compares to \( \beta \). A question with an imperfect \( \alpha \) may have a uniform probability distribution. Student performance on such a question gives no information with which to estimate \( \theta_s \).

Finally, a question with an anti-perfect \( \alpha \) is anti-correlated with \( \theta_s \), a situation that often means that grader is using a faulty answer key.

Conversely, if estimates of \( \theta_s \) are known it is possible to estimate the parameterization of each question. Repeated iterations of estimating \( \theta_s \) from question performance and then \( \alpha \) and \( \beta \) from \( \theta_s \) and answer scores is the traditional usage of IRT. Here we merely wish to leverage the concepts of difficulty and discrimination.

3. GOOD QUESTIONS

Given \( Q \) and an estimate of \( \theta \) for each student, we can evaluate individual questions using ideas and techniques from both information theory and IRT to quickly produce an \( \alpha \) and \( \beta \) estimate for each question. Based on these values, we can evaluate which questions are actually correlated with high ability in the course, and estimate the difficulty of each question. We can then begin to learn which types of questions are meaningful and which are not, as well as begin to get a sense of how difficult each question is. If the difficulty level for a question seems too high it indicates that the topic is not being understood by the students. Throughout this paper we will assume that a question with a high \( \alpha \) and a \( \beta \) within the meaningful range of \( \theta \) is a “good question.”

\(^1\)This is often reduced by characteristic curves that account for the probability that student will get the answer correct through random guessing.
3.1 Calculating $\alpha$ and $\beta$

Assume we know $\theta_i$ for every student for every latent trait that is tested in an instrument, and $\beta_j$ of the question. We can find the empirical discrimination for a question by evaluating how well that difficulty splits students with $q_{is} = 1$ from those with $q_{is} = 0$. That is, to what extent do students with $\theta < \beta$ have $q_{is} = 0$ while students with $\theta \geq \beta$ have $q_{is} = 1$?

To measure this, we can measure the entropy of $q_{is}$ for those students with $\theta < \beta$ and also for those with $\theta \geq \beta$. This gives a measure of the ability of the question to differentiate students at that $\beta$ threshold.

However, this is not quite sufficient, as a question that is perfectly anti-correlated (everyone with $\theta < \beta$ gets it right and vice-versa) will be given the same score as a question that is perfectly correlated. As such we do not simply want a measure of how homogeneous the two populations are. Hence we use a modified measure, not of how homogeneous the populations above and below the split point are, but rather of how well the split point puts everyone with $q_{is} = 0$ into $\theta < \beta$ and $q_{is} = 1$ into $\theta \geq \beta$.

$$\alpha = \frac{-\log_2 \frac{c_0(\theta < \beta)}{n}}{-\log_2 \frac{c_1(\theta \geq \beta)}{n}}$$

Additionally, a penalty term of $\Pr(\theta < \beta)$ is added if no incorrect answers fall on the low half of the split, and $\Pr(\theta \geq \beta)$ if no correct answers fall on the high half.

By making these changes we ensure that a perfectly anti-correlated question will be given a score of 1, even though the resulting populations are perfectly homogeneous. The above will give an answer in the range from 0 to 1. In order to make the score more intuitive, we reverse that range, so a score of 1 indicates a perfect discriminator, a score of 0 indicates a perfect anti-discriminator, and a score of 0.5 indicates that a split at that point gives no information about the true value of $\theta$.

Given such an evaluation function, here are a few conceptual examples to clarify.

- **Anti-correlation**: $\alpha = 0$ means that every student with $\theta \geq \beta$ answered the question incorrectly, and every student with $\theta < \beta$ answered the question correctly.

- **No correlation**: $\alpha = 0.5$ means that the students with $\theta \geq \beta$ answered correctly as often as incorrectly, as did students with $\theta < \beta$.

- **Perfect correlation**: $\alpha = 1$ means any given student got the correct answer if and only if she/he has an ability of $\theta \geq \beta$. In practice this means we are certain that this is the difficulty level for this question. Such scores will generally only be seen in small classes.

Given this ability to produce $\alpha$ from the distribution of $\theta_i$, and the vector $q_i$, we can find the best approximation of $\beta$ by looping through all meaningful split points on the $\theta$ axis and maximizing $\alpha$. In practice this means finding $\alpha$ for each distinct value of $\theta_i$ in the class.

3.2 Assumptions

In order for the above to be usable, several assumptions and simplifications must be made. Most instructors will not have estimates of $\theta_i$ for every student and topic that makes up a curriculum. Thus we must approximate the latent ability scores of the students. If we assume that the question is testing a topic related to the rest of the course$^2$, then we can use a value that is already being tracked: the student’s course score$^3$.

This assumption implies that each course as a whole is being treated as a single proficiency. This is clearly not always valid, but for any course for which this is a bad assumption (for example, a course composed of two half-courses on different topics), we can use $\theta$ values only from the appropriate portions of the course.

The estimates of both $\beta$ and $\alpha$ for each question rely heavily on the distribution of $\theta$ values. Student grades fluctuate significantly at the beginning of the course before the law of large numbers begins to stabilize each student’s grade toward its final value. Therefore, the values of $\beta$ and $\alpha$ are going to be most accurate at the end of the course, and a final evaluation of which questions are worth keeping for next quarter is best done after the course has completed.

It is useful to evaluate questions immediately after grading the instrument, but such evaluations should be recognized as estimates.

Independence of scores is also a concern. If the estimate of $\theta_i$ includes $q_{is}$, then a high $\theta_i$ by definition has some correlation to $q_{is} = 1$. To get the best estimates of $\alpha$ and $\beta$, it is best to provide $\theta_i$ such that the score on this question/instrument is excluded. In practice, if each individual question has a very small effect on the total grade, then this effect will be negligible and scores need not be recalculated. Tests with a large course-weight and few questions should be evaluated separately.

3.3 Sample Results

This technique of evaluating questions was used during our Introduction to Data Structures and Algorithms course during Summer 2004, where it provided very useful feedback to the instructor in evaluating the questions asked on three 10-Question quizzes, the 34-Question midterm, and the 34-Question final examination. Some of the intuitive things it confirmed included:

1. **Do not test on specifics from the text**: Questions that test not the subject, but merely a particular presentation of the subject have little assessment value. On Quiz 1, the second-worst $\alpha$ score was for a question of this type. Some instructors may choose explicitly to test this material to encourage students to do the required reading, but that was not the goal here.

2. **Do not test things mentioned in passing**: While some may feel that mentioning something once during lecture is sufficient for the “good” students, this appears not to be the case. Iterators were discussed by an aside for 10-15 minutes the day before Quiz 1. This was not nearly enough for even the good students to pick up the material. Although 38% of the class answered this

$^2$A counter example: if the topic of the class was Data Structures and a given question was about the instructor’s cat, then it is not necessarily the case that ability in the rest of the course correlates with likelihood of answering that question correctly.

$^3$This technique does not require any particular domain for these scores to be reported in, so percentages or raw scores are equally valid.
multiple choice question correctly, it is likely because they narrowed it down to 2 or 3 choices and guessed rather than knowing the answer, since this question had \( \alpha = 0.56 \), implying practically zero correlation.

The most useful feedback that this technique provided was in finding questions with unexpectedly high values of \( \beta \). For example, “What is the worst case time to find an element in a binary search tree of \( n \) nodes?” was asked on the second quiz. While there is a mild “trick” in recognizing that a binary search tree is not necessarily balanced, it was shocking to find that the \( \beta \) score for this question at that point was 89%, with an alpha of .72, meaning it did a better job of differentiating A’s from B’s than anything else, and generally did a fair job of that. This question was certainly not intended to be splitting the A’s from the B’s. The fact that it was found to be quite difficult indicated that the maximum depth of a binary search tree was a topic that needed additional coverage. Since the quiz was graded and evaluated the same day it was issued, the lesson plan for the next day was altered to account for the fact that the class had missed this rather critical concept.

This analysis technique provides immediate feedback in two important ways. First it allows instructors to determine which questions are good questions, and can be used to train instructors not to ask overly vague questions. Second, and more importantly, it can provide concise feedback on what students are actually understanding and what they are only guessing. In this way, topics that were confusingly or incompletely covered can be immediately brought to the instructor’s attention and remedied as quickly as possible. Both of these seem to be very valuable in our efforts to increase educational effectiveness.

4. GOOD INSTRUMENTS

A higher-level issue than “Was this a good question?” is the question of “Was this a good instrument?” Given the assumption that the content of the instrument is not orthogonal to the rest of the course, we can measure how effective and on-topic an instrument is using some basic concepts from information theory introduced in Section 2.

4.1 Instruments and Grades

We would like to calculate the correlation between student performance on the instrument and \( \theta_i \). This will tell us how much the instrument is measuring what we think it is measuring. This is the mutual information between \( \theta \) and the set of score vectors \( q_i \) (which together is simply \( Q \)).

\[
I(\theta; Q) = H(Q) + H(\theta) - H(\theta, Q).
\]

This presents a difficulty. We need to estimate the distribution of answers over the space of all questions. If the instrument has \( m \) binary (right-or-wrong, no partial credit) questions, then there are \( 2^m \) such possible events. In order to have reasonable data sizes to pursue this avenue, we would need a number of students \( n \) proportional to \( 2^m \). Without this, then the majority of possible student response vectors will be unseen and hence have an estimated probability of 0. It is certainly not the case that there is no possibility for a given set of answers on a test, so we are not comfortable ignoring this fact.

Several simplifying assumptions can be made to give an approximation of \( I(\theta; Q) \). The most useful of these based on our investigations is to calculate the empirical mutual information between the grade \( g \) and the total number correct on the instrument \( S = \sum_{i=1}^{m} q_i \). These both have a much smaller range of values, and have shown in our experiments to yield meaningful values on real data. We used the A/B/C/D/F grading system, although any discrete grading system would work.

This empirical mutual information can be expressed as

\[
\hat{I}(S; g) = \hat{H}(S) + \hat{H}(g) - \hat{H}(S, g).
\]

\( \hat{H}(g) \) has a known upper-bound proportional to the log of the number of buckets in the discretization (thus, \( \log 5 \) if A/B/C/D/F is used). \( \hat{H}(S) \) however is only bounded by \( \log_2 \min(m, n) \), since \( \min(m, n) \) is the number of possible different aggregate scores of \( n \) students on a binary-valued instrument of \( m \) questions. This means that two instruments are not directly comparable, as they do not have the same upper-bound. An instrument given to a larger class will on average get a higher mutual information score than one given to a smaller class, and similarly the mutual information will rise with the number of useful questions. This is suboptimal from a mathematical point of view. However, so long as the questions are useful and pertinent to the course, larger instruments really do a better job of assessing student knowledge than smaller instruments. Thus, even though both tests don’t necessarily have identical ranges for \( \hat{I} \), the notion of one being more valid than another does apply for instruments assessing similarly-sized classes.

When calculating \( \hat{I}(S; g) \) the only information necessary is information that is already being recorded, namely aggregate scores on each instrument and course grades. Unlike our assessment of individual questions which requires more information that might not be collected by instructors, this feedback mechanism can be utilized immediately by everyone.

4.2 Results

Based on the relatively small set of truly comparable (same number of students and questions) data sets available, this measure captures the essence of the detailed information available through question-level analysis.

One set of comparable instruments is three quizzes of 10 questions and 16 students each. The first quiz had six questions with \( \beta \) values less than 60%. As only one \( \theta_i \), was less than 60%, these questions are therefore not really contributing much information to the overall course grade, or the to meaningfulness of the instrument.

The second quiz had a more uniform distribution of difficulties, and the highest average \( \alpha \) of the three.

The third quiz had five questions with \( \beta \) values at or above 83%, while only 38% of the class had \( \theta \) scores at or above that level. Two of the questions had \( \beta \) at 92%, higher than all but one \( \theta_i \). If the mutual information measurement has any meaning, this quiz should score better than the first quiz, but not as high as the second.

As predicted, the first scores a 1.97, the third a 2.04 (better, but not significantly), and the second a 2.31. These numbers represent how much information is shared between measurements of \( g \) and \( S \), which is strongly related to \( \alpha \) being high and \( \beta \) being well distributed across the range of \( \theta \) system would work.

The results are presented primarily as a check on the validity of the empirical estimates of \( H(g) \) and \( H(S) \). The
concept of evaluating the effectiveness of an instrument is relatively foreign to most instructors. The goal, however, is primarily to provide some method of closing the feedback loop on instrument creation so that an instructor may, without spending a great deal of time, get some sense of how “good” the instrument was. It is our hope that this score will fulfill that requirement.

5. RESULTS
The “best” and “worst” of the three quizzes discussed in Section 4.2 are provided in Table 1 as an example of the effectiveness of these techniques. Also provided for comparison is the percentage of students that answered the question correctly.

In this table it is interesting to note that there is little correlation between average percentage correct and being a useful question. Especially on hard questions, the $\alpha$ value gives insight whereas the percentage correct gives little information about whether the question was hard or misleading. An example of this is Question 9 from Quiz 1: with 65% of students responding correctly, this question is obviously somewhat difficult, but not out of the ordinary. However, with an $\alpha$ of 0.56, the question is essentially telling us nothing: the top students are guessing just as much as the struggling students. Another example is Question 6 on Quiz 2, with $\alpha = 0.6$ and 71% answering correctly. From percentage correct alone, there is nothing to indicate that there was some uncertainty about the question. An $\alpha$ of only 0.6 on a True / False question is really quite telling: there was something missing for this question, either in the wording or lecture coverage. Throughout the table, $\alpha$ gives more useful information, and the combination of $\alpha$ and $\beta$ contains significantly more information than the percentage correct alone.

6. SOFTWARE AVAILABILITY
A Python script capable of performing the analysis detailed in this paper is available from [5] along with the instruments discussed and anonymous scores on those instruments. This software is provided under the GPL.

7. CONCLUSIONS
Every attempt on every question is fundamentally testing two things: how good is the student, and how good is the question. In the normal course of teaching, we must evaluate students continuously, and with the aid of the techniques introduced here we can begin to evaluate the questions as well. Evaluating educational effectiveness is an important concept in current accreditation practices, and we expect that importance only to grow in years to come. The techniques presented in this paper attempt to close the loop on two of the lowest-level feedback processes by reinforcing which questions and which instruments are the most meaningful with respect to the course grade, and are certainly more informative than the simple “percentage correct” statistic that is sometimes used as a check on item effectiveness. In short, we have developed two simple estimates to provide a low-effort, high-payoff solution to evaluate the effectiveness of testing instruments.

<table>
<thead>
<tr>
<th>Statement</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quiz 1: M.I. Score: 1.98. Average %: 78.8</td>
<td>0.94</td>
<td>0.52</td>
<td>1</td>
</tr>
<tr>
<td>A stack is a FIFO</td>
<td>0.94</td>
<td>0.52</td>
<td>1</td>
</tr>
<tr>
<td>A stack built using ADT list adds and removes from different ends of the list</td>
<td>0.78</td>
<td>0.52</td>
<td>0.88</td>
</tr>
<tr>
<td>A stack using a linked-list can be implemented so push and pop are $O(1)$</td>
<td>0.84</td>
<td>0.59</td>
<td>0.76</td>
</tr>
<tr>
<td>A queue is a FIFO</td>
<td>0.94</td>
<td>0.52</td>
<td>1</td>
</tr>
<tr>
<td>A queue can be implemented as a singly-linked list so enqueue and dequeue are $O(1)$</td>
<td>0.84</td>
<td>0.59</td>
<td>0.76</td>
</tr>
<tr>
<td>You can only write an iterator for a linked list</td>
<td>0.78</td>
<td>0.59</td>
<td>0.82</td>
</tr>
<tr>
<td>For ADT List, which of these is not part of the basic set of operations?</td>
<td>0.66</td>
<td>0.76</td>
<td>0.65</td>
</tr>
<tr>
<td>A linked-list can never</td>
<td>0.84</td>
<td>0.74</td>
<td>0.82</td>
</tr>
<tr>
<td>Which of the following statements about Iterators is not true?</td>
<td>0.56</td>
<td>0.93</td>
<td>0.65</td>
</tr>
<tr>
<td>An uncaught exception (one that matches no catch blocks) will:</td>
<td>0.77</td>
<td>0.67</td>
<td>0.65</td>
</tr>
</tbody>
</table>

| Quiz 2: M.I. Score: 2.31. Average %: 70.6                                 | 0.86  | 0.52  | 0.94 |
| A sorting algorithm accessing a list in sequential order can be done on a linked-list with an iterator. | 0.78  | 0.52  | 0.88 |
| The following is code for which sort?                                    | 0.62  | 0.69  | 0.71 |
| Pick an element of the list as a “pivot”                                 | 0.71  | 0.93  | 0.29 |
| Move everything less than the pivot to one side and everything greater to the other and recursively sort both sides. Which sort is this? | 0.72  | 0.87  | 0.29 |
| The following code will do what?                                         | 0.60  | 0.80  | 0.71 |
| Since Radix Sort is $O(n)$ and Bubble Sort is $O(n^2)$, Radix Sort is ALWAYS faster. | 0.77  | 0.52  | 0.76 |
| Quick sort has a worst case runtime of $O(n \log n)$                     | 0.72  | 0.80  | 0.65 |
| Merge Sort cannot be used efficiently in place                          | 0.78  | 0.76  | 0.82 |
| The proper evaluation of this postfix expression is what?               | 0.94  | 0.52  | 1   |

Table 1: Analysis of Two Quizzes

8. REFERENCES