Proceedings of the Twenty-Ninth International
Florida Artificial Intelligence Research Society Conference

Comparative Methods and Analysis for Creating High-Quality Question Sets from Crowdsourced Data

Sarah K. K. Luger and Jeff Bowles
Institute for Language, Cognition and Computation
University of Edinburgh
Edinburgh, UK
s.k.k.luger@sms.ed.ac.uk; jbowles@cs.unm.edu

Abstract

Online assessment has grown beyond the confines of conventional educational testing companies. Assistance in creating high-quality exams is welcomed by educators who do not have direct access to the proprietary data and methods used by educational testing companies. Using aspects of accepted educational approaches to measure question difficulty and discrimination power, this paper covers two methods for building exams composed of high-quality multiple choice questions (MCQs) from sets of crowdsourced data.

Introduction

Traditional MCQ exams are built by educational testing companies with algorithms and data sets that are unavailable to the general public. We present a method that efficiently builds new exams using crowdsourced data.

First we present a graph-based representation for gathering training data from existing, web-based resources that increases access to such data and better directs the development of good questions. Then, we present a complementary method based on weighting questions by difficulty for building an exam. Further, using Item Analysis Theory, (Gronlund 1981), we analyze these new virtual exams and measure both the item difficulty and the discriminating power of the questions. These measures suggest characteristics that can be used by an automated question analysis system for rating question difficulty and discrimination.

Then, we present a method that efficiently builds new exams that consist of only these discriminating questions and we demonstrate the effectiveness of this new question set by monitoring student performance group movement across exams of different sizes. This supports the determination of an optimal exam size and question difficulty-level to achieve maximum subject discrimination.

The data used in this work is from the PeerWise question authoring and answering community (Denny 2009). Using the PeerWise interface students create MCQs and other students answer and refine these questions. The questions used in this research are from two courses of a university-level introductory biology class.

Crowdsourcing MCQs and Item Analysis

The MCQ in Example 1 is from PeerWise and was answered 133 times, 86 correctly.

Example 1: What is the sulcus/fissure called that divides the brain frontal from parietal?

A. Lateral Fissure
B. Parietoccipital Sulcus
C. Longitudinal Fissure
D. Central Sulcus *correct answer*

To measure the usefulness of exam questions, researchers have devised methods for judging both the difficulty of the question and the differentiation power of the answer options (Patz and Junker 1999) and (Beguin and Glas 2001). Probabilistic graphical algorithms that model the difficulty of questions and the quality of test-takers have been tested on crowdsourced data (Bachrach et al. 2012a) including in situations where there is no correct response information (Bachrach et al. 2012b). Item Response Theory has been previously used to evaluate the quality of questions in crowdsourced tasks (Christoforaki and Ipeirotis 2014). In the closely related Item Analysis Theory a group (for example, 100 students) takes a test containing suitable questions and the exams are graded and ranked. Then, the set of 100 students is split into three cohorts, that represent the top-, middle-, and lowest-scoring students. These three cohorts are defined as the lower 27% and top 27%; the

middle 46% is excluded because it offers little meaningful information. The process for performing Item Analysis is:

1. For each test item (question), the number of students in the upper and lower groups who chose each answer option is tabulated in a template.
2. Item Difficulty is measured by the percentage of students who answered a question correctly. The lower the percentage, the more difficult the question.
3. Item Discriminating Power is the difference between the number of high-scoring students versus the number of low-scoring students who chose the same answer option.

This method for judging question difficulty and item discriminating power relies on three cohort-based models of student performance. Comprehension and aptitude tests seek to present questions that can be correctly answered by students who understand the subject matter and to confuse all other students with seemingly viable alternate answer options (distractors). A *good* or difficult distractor is one that catches or distracts more bad students than good students; such items have a positive number in the "Diff" column in the Item Analysis examples.

A high-scoring student is one who answers most questions correctly, but when their answers are incorrect, chooses the best distractors. A low-scoring student will choose any of the answer options seemingly at random. A difficult question is one whose answer options are all deemed viable to a high-scoring student. With a difficult question, the high-scoring cohort will behave like low-scoring students, with a near equal spread of multiple distractors being chosen. In summary, we measure the relationship between the question and the answer option following Item Analysis' perspective on student performance.

The Two Exam-Building Methodologies

An exam is a set of students who have answered the same questions. This is the scenario that one would see in a conventional school classroom with the students taking a test. The PeerWise data sets consist of students who have answered some questions, but not necessarily the same questions from a set. Thus, the data contains an incomplete, or sparse exam. In the experimental question sets there are:

Course:	1	2
Total number of students:	1055	887
Total number of questions:	148	132
Shared edges between the questions and the students:	28049	31314

We present two approaches for turning this valuable, but sparse data into useful exams. One approach is to find those questions that most students answered in common. We need to include the same students who have answered the same questions because we are attempting to use Item Analysis that is dependent on full exam-based results. This allows us to discover information after the entire exam has been graded on question difficulty and student performance.

Another approach is to choose the questions based on difficulty. A good exam should only include questions that are moderately difficult. This difficulty-based approach eschews exam-based Item Analysis and provides an interesting comparison for the clique-focused method.

Clique-based Methodology

Our approach for representing the individual student question answering relationship is with a graph: an "exam", where every student answers every question would be a complete bipartite graph (or biclique). We are seeking a good set that is similar to an exam.

Consider this problem of finding the sufficiently large set of the same students and the same questions in an exam as that of comparing edges in a graph, where the goal is to find an edge between S1, student 1, and Q1, question 1. Next, searching the edges of the nearby students and questions may provide another student who has answered the same question as the S1→Q1 pair. Thus, every student is compared to the first student, S1, to check if they have also answered the first question. This iterative checking of group membership is exacerbated by the possibility that the first student whose answered questions were compared against all other questions in the set did not answer all of the questions. As a result, once all of the questions that S1 has answered have been compared to all of the answered questions of all of the other students, starting a new search with S2 may provide either more or fewer shared questions with the set of all other students. See Luger and Bowles 2013 for a deeper clique-based methodology discussion.

In creating virtual exams out of sets of questions, the aim is to discover sets that satisfy two parameters (the most same students answering the most same questions). Even though this is an NP-hard search problem, due to the fact that some resulting virtual exams are *better* than others, it is also an NP-hard optimization problem. In one case, what makes one set of students and questions *better* than another similar set is based on what balance of students and questions provides an exam that allows for optimal Item Analysis. The second feature that makes a *better* virtual exam is the balance of discriminating questions. In other words, an exam with 10 discriminating questions and 20 students is superior to a larger exam that had fewer discriminating questions. We seek discriminating exam questions.

Given an incidence matrix M of students and questions, where the rows of M correspond to students and the col-

umns correspond to questions, we can generate covariance matrices S and Q. S is defined as $M \times M^T$ which generates a covariance matrix where S_{ij} shows how many questions student i has answered in common with student j. Q is defined as $M^T \times M$ which generates a covariance matrix where Q_{ij} is the number of times questions i and j were answered together. S and Q can then be used heuristically to compute a sufficiently large clique of questions that have all been answered by the same set of students.

Steps for building and sorting the covariance matrices:

1. Collect the data in triples of student ID, question ID, and answer choice. The students are ordered by the number of questions they answered.
2. Build the incidence matrix M with students listed as rows and the questions as columns. The incidence matrix can also be expressed as a bipartite graph.
3. Compute $S = M \times M^T$ and compute $Q = M^T \times M$.
4. We can find the most correlated students by computing the vector **s** by summing over the rows of S. Thus $\boldsymbol{s} = \Sigma_i S_{ij}$. We then sort the rows and columns of S based on the ordering of **s** as S is symmetric.
5. As above, we can find the most correlated questions by computing the vector $\mathbf{q} = \Sigma_i Q_{ij}$. We then sort the rows and columns of Q based on the ordering of **q**.

This sorting process provides a sound heuristic for selecting highly correlated students and questions. We then selected the top 15% most correlated students and the top 15% most correlated questions from the dense group of students who have answered the same questions. This presents a realistic exam where there are a few holes, i.e., omitted questions. These are missing edges in a bipartite graph. Once we have built the exam, we analyze the individual question difficulty and discrimination power. In this research, we followed the steps presented twice, once for each of the two courses of students and questions.

Question Weighting-based Methodology

This method takes a different approach than the clique-based method that uses Item Analysis, builds incidence matrices, and finds highly correlated questions and students. The second method weights the individual questions based on how every student that tried the question performed. The students are scored based on how all of the students who tried each question answered. Since this approach does not depend on creating sets of correlated questions and students, it contrasts with the clique-based approach of turning sparse student-question matrices into denser exam data for scoring exams. Nonetheless, the goal of the weighting approach remains the same: find the least discriminating questions and remove them from the set.

In this method, the assumption is that questions that are very easy or very difficult are not discriminating. Questions that are too easy or too hard do not reveal any information about the student in comparison to their peers. Thus, these questions do not discriminate. Meaningful information about how students perform is relayed by questions that only the high- or only the high- and the middle-performing students answer correctly. Questions that all of the students get incorrect or all of the students get correct do not tease apart the variations in the comprehension levels within the larger group of students.

Finding the boundaries between too easy, too hard, and discriminating is an iterative process where questions at either end of the list are removed one-by-one and students are scored based on the remaining questions and the question weights. The resulting student score represents how a student performed on the questions that they attempted. Since not all of the students answered the same questions, this helps to differentiate a better performing student from a lower performing one if both students answered all of the questions they tried correctly, but one student attempted much more difficult questions. Students who perform well on harder questions would be rated as better than those that perform best on easier questions. We are trying to correct for question difficulty level self-selection without using the exam model in the clique-based approach that incorporates making the students answer the same questions.

To calculate the weights of the questions, a list of students is created who have answered at least three questions. Three questions is the minimum needed to separate a group of students into performance cohorts (high-, middle-, and low-performing students). A "weight" vector, **w** is created where each element of the vector is the weight for a question. The questions are weighted based on the number of times a question was answered correctly. Weights are normalized, or in the range [0, 1]. A question with weight 0 is a question that was never answered correctly; a weight of 1 is given to a question that was always answered correctly. Components of the weight vector are calculated using:

Where x is the position in the vector **w**, $n(x)$ is the number of answers to question x, $c(i,x)$ is the correctness of student i's answer to question x. Values for $c(i,x)$ are 1 if the answer is correct, 0 if wrong. Weights are in the

$$w(x) = \frac{\sum_{i=0}^{n} c(i,x)}{n(x)}$$

range [0,1] where weights closer to 0 correspond to very difficult questions and weights closer to 1 correspond to very easy questions. We aim to find the set of discriminating questions that are neither too easy, nor too difficult.

In the two courses, the questions are of moderate to easy difficulty. A few of the hard questions were answered correctly by about 1 in 5 students, but the majority of questions were answered correctly by more than 1 in 2 students.

Application of Methodologies

Now that we have presented how the clique-based and weight-based approaches work, we will step through them in the following section. A discussion comparing the merits of the two approaches is included in the final section.

Applying the Clique-based Methodology

All of the steps in exam building to this point are performed to test the individual questions in an exam for their difficulty and discriminating power. The first goal is to ask questions at the correct level of difficulty that successfully judges comprehension. Question difficulty is an important measure that can help direct how an exam is built including helping an instructor gauge the intended level of difficulty. The second goal is to ask questions that group the students into performance cohorts.

Question difficulty is linked to question discriminating power. Question difficulty is measured by the number of students who answered a question correctly divided by the number of students taking the exam. As a rule, the lower the percentage difficulty, the harder the question. If a question is too hard, all of the students will answer incorrectly, and if it is too easy, they will all get it correct. If a student omits a question, that counts as an incorrect answer.

Question discrimination looks at not only whether a student got an answer correct, but if their behavior –the distractor they chose– mirrors the behavior of other students. These other students can be grouped into performance cohorts that are used to measure how good a question was at having the high-performing students answer it correctly, but not being answered by the lower-performing students. Thus, we also can determine the most discriminating questions in an exam. Discriminating questions most effectively sort students into their relative cohort, and are the most valuable questions in exams as they provide the most information about a student's level of comprehension.

Example 1 had a difficulty of 65% and Example 2 had a difficulty of 60%. Example 2 is a non-discriminating question because answer B, the correct answer, was chosen by a large majority of the high-performing students as well as the students in all other groups. It failed to sort students.

Example 2: Which is an afferent pathway in which the axon cross over immediately after entering the spinal cord?

A. Medial Lemniscal
B. Anterolateral *correct answer*
C. Corticospinal
D. Noncorticospinal
E. Nonmedial

(Low/negative discrimination power, omission rate .038)

The reader may notice that the "Omit" row, "Diff" column entry is always zero, because this value is never calculated as it has no bearing on the question's Discrimination Power. Example 2 was initially answered 282 times, 155 correctly. In this exam based on 158 students (the top 15% most correlated students), the three cohorts are shown on the row titled "Total". The cohorts are 35 good, 62 average, and 36 low-achieving students.

Letter	High	Middle	Low	Total	Diff	
A	3	7	6	16	-3	
B	25	41	16	82	9	Correct
C	2	2	1	5	1	
D	0	3	3	6	-3	
E	5	7	6	18	-1	
OMIT	0	2	4	6	0	
TOTAL	35	62	36	133	0	

Example1: See Crowdsourcing MCQs for question stem and answer options. Option D is correct.

(High/positive discrimination power, omission rate .053)

Example 1 is a discriminating question because it has multiple strong answer options that distract more high-performing than lower-performing students. The "Diff" column takes the difference between what answer options the high- and low-scoring students choose and produces a zero, a negative, or a positive number. Zero equates to a question having no discriminating power. A negative number means that lower-scoring students were more attracted to this option than their higher performing peers. Positive numbers in the "Diff" column reflect a *good* distractor. A more difficult question usually has more good distractors.

Letter	High	Middle	Low	Total	Diff
A	8	11	7	26	1
B	2	3	5	10	-3
C	2	1	1	4	1
D	23	46	17	86	6

The Question Weighting-based Methodology

After the questions' weights are calculated, the questions are sorted based on their weight. Students are scored by taking the sum of all of the question weights for the correctly answered questions, and then dividing by the sum of all the questions weights for the answered questions.

$$s(i) = \frac{\sum_{x=0}^{m} c(i,x)w(i)}{\sum_{x=0}^{m} a(i,x)w(i)}$$

Where m is the number of questions that a student has attempted to answer, and $c(i,x)$ is the correctness of student i's answer to question x. Values for $c(i,x)$ are 1 if the answer is correct, 0 if wrong. $a(i,x)$ is 1 if the question was attempted by student x, 0 if not. The denominator has the effect of normalizing student scores into the [0,1] range.

After scoring, the students are rank-ordered and placed into three cohorts: high-, middle-, and low-scoring students. The size of the cohorts remain constant and are split into lower 27%-middle 46%-upper 27%. The rank-ordered

list of questions represents a set of weights and is also referred to as the "spectrum". This methodology repeats the formula for building performance cohorts.

Next, we seek the set of the most discriminating questions. Questions at either end of the weight spectrum are removed one-by-one in an effort to find the central band of discriminating questions. This process consists of removing a single question from one end of the spectrum, scoring the students and placing each student into a cohort, which is equivalent to building a histogram of students based on their scores. In essence, the goal of this process is to remove questions from the possibly non-discriminating question list while limiting changes to the histogram. We perform this operation by sorting question weights and under the assumption that the more discriminating questions are near the middle of the spectrum and that the less discriminating questions are closer to either end of the spectrum.

This process is applied repeatedly and cohort movement is measured. In the scoring process, students are not penalized for questions that were not answered, they are only penalized for wrong answers. This process outputs exams, student scores, their cohort, and the amount of cohort movement from the initial exam that includes all questions to the current exam. To compare with the clique-based method, exams of the same size are created and cohort movement is measured for both methods. This is done to ensure that cohort movement can be compared equally between both methods.

Results and Discussion

In this section, results from both the clique-based and question weighting-based methodologies are discussed. We show the optimal parameters for applying the clique-based methodology and describe the results. We also show how the question weighting-based methodology provides lower quality results. Finally, we discuss observations from analyzing cohort movement while creating useful exams.

Clique Methodology

Student cohort movement occurs as students answer each additional question. For example, three students begin in the same cohort and as they answer a question either correctly or incorrectly, their choices place them into different performance groups (high, middle, and low). Each performance group is distinctly differentiated depending on how students perform and the movement may suggest reordering questions to show and stabilize the cohorts more quickly. Cohort movement occurs when a student is in one performance group and then their responses sort them into another cohort. We have found that student movement between performance cohorts is quite high, around 30% when there are very few questions in the exam.

Desiring stability in performance cohort movement motivated using 15-30% of both correlated students and questions. Comparing the balance of the number of students and the number of questions in the exams suggested using the most correlated 15% of students and the most correlated 25% of questions for further analysis. After we created the exams from the most correlated 15% of the students and the most correlated 25% of the questions, we used Item Analysis to find the most difficult and most discriminating questions within the exams.

Question Weighting Methodology

The question weighting methodology produced very different results when compared to the clique-based methodology. When questions with low and high weights were removed from the list to find exam sizes that were the same as the clique-based methodology, we found that 44% and 46% of the students were scored so significantly differently that they would be moved into different cohorts. In contrast, in the clique-based method only 11% and 20% of the students moved into a different cohort. This indicates that performing analysis based on question weights is not an attractive method for finding the most discriminating questions and that Item Analysis combined with our clique-based methodology provides a more robust solution.

Question weighting was viewed as an alternate method for finding the most discriminating questions, but it appears that this analysis does not take into account enough contextual data to discover the most discriminating questions. It was assumed that questions with low and high weights would not have much discriminating power, but when this method was applied, cohort movement was unacceptably high.

In the question weighting method, students who answered fewer than 3 questions were omitted from the analysis. Their answers were not considered in the weighting, nor were they scored as part of the cohort measurement process. This resulted in a reduction of 16% and 10% of the students considered in courses 1 and 2, respectively. In course 1, over 50% of the students scored 75% or better. In course 2, 44% of the students scored 75% or better. This perhaps indicates better students answered more questions and will be examined in future work.

Steps Towards Creating the Ideal Exam

After completing the MCQ analysis, we can now ask which and how many questions are needed to create a quality exam. We presented two methodologies that filter out the least discriminating questions in an exam. The first approach analyzes the best balance of students and questions based on creating a more dense matrix of those students and questions. The second approach analyzes the

questions' difficulty to find the best new exam set that includes the most discriminating questions.

We began in the first course with a sparse matrix of 1055 students and 148 questions and using the adjacency matrix approach created a new exam comprised of 158 students and 37 questions. Once we discovered a sufficient new exam size we performed Item Analysis and measured the discriminating power of each question. The filtering is based on a question having a net positive Discriminating Power. After correcting for Item Discriminating Power, the new exam size was 26 questions answered by 158 students.

The second course began as a sparse matrix of 887 students and 132 questions and using the matrix approach became a new exam comprised of 158 students and 37 questions. After correcting for Item Discriminating Power, the new exam size was 20 questions answered by 133 students.

In course 1, the new exam cohort movement was low at 11%. 18 students moved from one performance cohort to another based on how they answered the first set of questions as compared to the second set. In course 2, the cohort movement was 20%. In both courses, the first set of questions was the top 15% most correlated students and the top 25% most correlated questions. In the second set, the questions deemed the least discriminating were filtered out. Note in Table 1 that cohort movement only occurs from the cohorts next to one another, such as the middle to high.

Method:	Clique	Clique	W	W
Course:	1	2	1	2
Total no. S:	1055	887	886	807
Total no. Q:	148	132	148	132
% top correlated S:	0.15	0.15	N/A	N/A
% top correlated Q:	0.25	0.25	N/A	N/A
Omissions:	YES	YES	N/A	N/A
New no. S:	158	133	886	807
Initial exam size:	37	32	148	132
New exam size:	26	20	26	20
Cohort movement:				
Low to middle:	4	5	43	25
Low to high:	0	0	29	24
Middle to low:	4	5	125	121
Middle to high:	5	8	70	72
High to low:	0	0	22	27
High to middle:	5	8	105	106
Numerical total:	18	26	393	375
% total:	0.11	0.2	0.44	0.46

*Table 1: Comparing the clique-based and weighted (**w**) methods shows that cohort movement only occurs in neighboring cohorts.*

Conclusion and Future Research

We have addressed the task of making MCQ exams by considering crowdsourced data. We repurposed exam preparation materials to generate new exams and discover exam questions that are both difficult and differentiating.

We demonstrated two sets of algorithms that identified appropriate MCQs from PeerWise and showed how these questions could be analyzed to determine both their difficulty and discrimination. First, we presented our matrix-based method for data analysis and then built exams out of sets of questions that have been answered by students. Second, we looked at the questions and weighted them based on the performance of all students who attempted them.

We then used that difficulty measure to weight the performance of the students and created a ranked spectrum. This approach, created less stable performance cohorts, is less viable than the clique-based exam-building method.

Future work includes analyzing other methods for building exams, especially ones that contain few omitted questions and cover all of the topics in a subject area. Discovering the ideal size of an exam for discriminating students into performance cohorts remains an open research area. Additional work will compare these results to those from supervised learning methods such as maximum likelihood estimation (Raykar et al. 2010).

References

Bachrach, Y., Graepel, T., Kasneci, G., Kosinski, M., and Van Gael, J. 2012. Crowd IQ - Aggregating Opinions to Boost Performance. *AAMAS 2012*.

Bachrach, Y., Minka, T., Guiver, J., and Graepel, T. 2012. How To Grade a Test Without Knowing the Answers - A Bayesian Graphical Model for Adaptive Crowdsourcing and Aptitude Testing. *ICML*, 2012.

Beguin, A. A., and Glas, C. 2001. Mcmc estimation and some model-fit analysis of multidimensional irt models. In *Psychometrika, Vol. 66, No. 4, pp. 541-562*.

Christoforaki, M., and Ipeirotis, P. 2014. STEP: A Scalable Testing and Evaluation Platform. Second AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2014).

Denny, P. 2009. Peerwise. http://peerwise.cs.auckland.ac.nz/.

Feller, W. 1950. *An Introduction to Probability Theory and Its Applications*. John Wiley and Sons, Inc.

Gronlund, N. E. 1981. *Measurement and Evaluation in Teaching*. Macmillan, 4 edition.

Luger, S. and Bowles, J. 2013. An analysis of question quality and user performance in crowdsourced exams. In *Proceedings of the 2013 Workshop on Data-Driven User Behavioral Modelling and Mining from Social Media at CIKM 2013. pp. 29-32.*

Mitkov, R., Ha, L. A., and Karamanis, N. 2006. A computer-aided environment for generating multiple choice test items. In *Natural Language Engineering 12(2) pp. 177-194*.

Patz, R. J., and Junker, B. W. 1999. Applications and extensions of mcmc in irt: Multiple item types, missing data, and rated responses. In *Journal of Educational and Behavioral Statistics, Vol. 24, No. 4 (Winter, 1999), pp. 342-366*.

Raykar, V.C.; Yu, S.; Zhao, L.H.; Valadez, G.H.; Florin, C.; Bogoni, L.; Moy, L. 2010. Learning from crowds. *Journal of Machine Learning Research* Vol. 11 pp. 1297-1322.