THE UNIVERSITY
of EDINBURGH

Algorithms for Assessing the Quality and Difficulty of Multiple Choice Exam Questions

Sarah Luger

Doctor of Philosophy

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh

2016

Abstract

Multiple Choice Questions (MCQs) have long been the backbone of standardized testing in academia and industry. Correspondingly, there is a constant need for the authors of MCQs to write and refine new questions for new versions of standardized tests as well as to support measuring performance in the emerging massive open online courses, (MOOCs). Research that explores what makes a question difficult, or what questions distinguish higher-performing students from lower-performing students can aid in the creation of the next generation of teaching and evaluation tools.

In the automated MCQ answering component of this thesis, algorithms query for definitions of scientific terms, process the returned web results, and compare the returned definitions to the original definition in the MCQ. This automated method for answering questions is then augmented with a model, based on human performance data from crowdsourced question sets, for analysis of question difficulty as well as the discrimination power of the non-answer alternatives. The crowdsourced question sets come from PeerWise, an open source online college-level question authoring and answering environment.

The goal of this research is to create an automated method to both answer and assesses the difficulty of multiple choice inverse definition questions in the domain of introductory biology. The results of this work suggest that human-authored question banks provide useful data for building gold standard human performance models. The methodology for building these performance models has value in other domains that test the difficulty of questions and the quality of the exam takers.

Acknowledgements

I am very grateful for the tremendous help I have received during my research. I would like thank Prof. Bonnie Webber and Prof. Johanna Moore for sheparding me through this process. I would like to thank my office mates Dr. Annette Leonhard-MacDonald and R. Alexander Milowski for their support. My friends and colleagues Vasilis Karaiskos, Dr. Simone Teufel, Dr. David Talbot, Dr. Mirella Lapata, Sarah Burgundy, Karyn Johnson and Jeff Bowles were all instrumental in this work. Prof. George F. Luger and Raymond Yuen are also deserving of special thanks for their wisdom and patience. Paul Denny, the creator of PeerWise was especially generous with his data and without his help this research would not have been possible. Finally, I would like thank Dr. Claire Grover and and Prof. Henry S. Thompson for their essential insights.

Thank you all very much; I am most appreciative!

Lay Summary of Thesis

THE UNIVERSITY of EDINBURGH

The lay summary is a brief summary intended to facilitate knowledge transfer and enhance accessibility, therefore the language used should be non-technical and suitable for a general audience. (See the Degree Regulations and Programmes of Study, General Postgraduate Degree Programme Regulations. These regulations are available via: http://www.drps.ed.ac.uk/.)

Name of student:	Sarah Luger	UUN	S0231607
University email:	s0231607@sms.ed.ac.uk		
Degree sought:	PhD	No. of words in the main text of thesis:	44113
Title of thesis:	Algorithms for Assessing the Quality and Difficulty of Multiple Choice Exam Questions		

Insert the lay summary text here - the space will expand as you type.

Click here to enter text.
The focus of this work is to leverage existing sets of multiple choice questions that have been created by engaged students in several college classes for further experiments on what makes a multiple choice question difficult and what makes a student good. Using crowdsourced data for this research is novel, but is supported by extensive related work. The value of improving multiple choice question creation and evaluation is motivated by these questions being the predominant style in standardized student evaluation. A series of experiments are run that look at how different types of students perform in three general performance cohorts. Then I seek ways to automatically discern good students and discriminating questions, using matrix-based mathematics to group similar behaving students and similar difficulty of questions. The results are significant and show promise over existing educational theory approaches.

Document control

Related policies/regulations: www.docs.sasg.ed.ac.uk/AcademicServices/Regulations/PGR_AssessmentRegulations.pdf	
If you require this document in an alternative format please email Academic.Services@ed.ac.uk or telephone 0131 650 2138.	Date last reviewed: 15.05.15

K:\AAPS\D-AcademicAdministration\02-CodesOfPractice,Guidelines&Regulations\24-MainReferencesCopiesPolicies\01-Current\Assessment BOE SCC & Feedback\Forms\ThesisLaySummary

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Sarah Luger*)

Table of Contents

List of Figures

List of Tables

Chapter 1

The Analysis of Multiple Choice Question Exams

The goal of this research project is to create an automated method to both answer and assesses the difficulty of multiple choice inverse definition questions in the domain of introductory level biology. First, I automate the answering of these questions with a system that uses web queries to gather possible answers. Then I compare the answers of the automated system to the text of these questions. Second, I use a set of human answered MCQs to study question quality. To do this I analyze the difficulty of questions and the differentiation power of their answer options. The goal of this work is to find automated methods that aid in the creation and evaluation of multiple choice tests.

Before going deeper into the details of this research I will describe briefly here how it evolved. I started by taking 1000 multiple choice questions from introductory college level psychology and biology curricula. I then developed a solution pipeline based on the ROUGE bag-of-words matching metric. Besides having a goal of solving nearly twice the number of questions as random answering would, I wanted to analyze properties of the individual questions, including their "difficulty" and power of discriminating good from poor students. In the initial stages of research I only had correct answers to focus on – no real human test-taking data. I also realized my solution pipeline could be improved by using more sophisticated bag-of-words-based technology including WordNet.

My second approach relied on human multiple choice question answering data supplied by PeerWise, an open-source platform that contains useful data for this research which allowed me to analyze approximately 1900 students answering 300 questions. The domain was strictly university-level introductory biology. I improved my solu-

tion pipeline by using Lucene with its multiple indexing and comparison methods, as well as using WordNet. The combination of automated solutions with human use data supported my analysis of question difficulty and discrimination.

My results suggest that algorithms that can judge question quality can benefit both educators, who spend large amounts of time creating novel questions, and students, who spend a great deal of time taking tests. The current approach for measuring question difficulty relies on inspecting exam results (after the exams are taken) and examining the answer distractors that high-scoring students pick most often compared to those that low-scoring students chose. This method relies on an educational psychology-based model called Item Analysis [2], which looks at how good pupils perform and contrasts that with their lower-performing peers.

Finally, these findings may be exploited in other domains. The initial data set contains Multiple Choice Questions (MCQs) from the area of college-level, introductory biology, but I extend improvements made in assessing these questions' difficulty so that they can be used as a methodology both for students taking these standardized tests and for the companies producing them, such as the Regents and Advanced Placement Exams.

To inspect how answer option variance contributes to question difficulty, I consider MCQs from the biology domain. The test questions are directed toward students in an introductory-level biology college course and are of a type that present a definition of a key biological concept or process and then seek the name of that concept or process. These questions are called Inverse Definition Questions (IDQs) and are very popular in multiple choice exams in the sciences where familiarity with domain-specific nomenclature corresponds to a basic understanding of the domain itself.

I present an alternative method for building exams from sets of questions that students have answered. An exam is a set of questions that have been constrained to be answered by a group of students. Creating "new exams" based on existing data sets of exam questions answered by *some* students fills a current gap in testing development. Measures used to judge question difficulty can be applied to these questions and the characteristics that make them difficult can be modeled and applied to other questions that have yet to be given to a group of students. The analysis of existing exam data is crucial for measuring the difficulty of individual questions and for investigating the traits that make a question discriminate among cohorts of students.

The two groups who benefit most from improving the ease of creating high-quality questions are educators and students. Educators currently depend heavily on stan-

dardized tests and measures that professional educational testing companies provide to them at a cost. The approach presented in this dissertation allows an educator who has a group of questions that have been answered by some students to know if the questions are easy or difficult. Further, my approach can demonstrate whether these questions have in fact discriminated the better performing students from their lower-performing peers, which is, in fact, the goal of a good question. The ability to use existing data in the form of questions answered by some students reduces educators' dependence on professional, and largely closed-source, testing companies.

From the students' perspective, the difficulty and discrimination analysis I present has a benefit as it allows students to take fewer exams or at least take exams that contain fewer questions. The "teach-to-test" phenomena has swept schools that are constantly seeking information on the performance level of their students. Being able to reuse existing open source data frees students from the rigidity and lack of opacity associated with conventional proprietary tests. It also requires less time for the actual exam-taking as well as more understanding of the results of previous tests. Finally, the environment that allowed this research to proceed, PeerWise, shows promise to provide an avenue for question-authoring and self-testing, both which benefit the student, as will be discussed more in Chapter 2.

1.1 What Makes a Multiple Choice Question Difficult?

In quizzes, both recreational and compulsory, statements that succinctly describe a concept and ask for the name of that concept are a well-accepted method of testing the recall in a subject area [6] [7], and are referred to as Inverse Definition Questions (IDQs). If such a question does not present answer options, it can also be described as an inverse definition "slot-filling" question, because the tester seeks the missing name that the question describes. Often, these types of questions are used in assessing students' academic progress, which is linked to comprehending concept sets in various domains. According to Phelps [8], multiple choice versions of IDQs are prevalent in secondary school exams because they are an efficient and effective way to assess understanding of curricula.

A simple example of an IDQ is

> **Example 1**
>
> In the QA community, questions that present definitions and ask for the term being defined or ask for a word or phrase that refers to the entity or

process being defined, are known as

A. Inverse Definition
B. Inverted Descriptive
C. Quiz-Style Questions
D. **All of the Above (correct answer)**

Definition questions are questions that present a term and ask for relevant information that describe that term. Questions that do the inverse, i.e., that present definitions and ask for the name of the term being described are known as Inverse Definition, Inverted Descriptive, or quiz-style questions in the Question Answering (QA) community [6] [7]. IDQs are not dealt with in the Text REtrieval Conference (TREC), which leads QA research. Moreover, they have been identified as not only "difficult to answer," but also requiring "urgent" attention because they are so inadequately answered by conventional QA systems [7].

For example, if a definition question is: "What is a phenotype?" then one possible Multiple Choice IDQ could be

Example 2

The outward appearance (gene expression) of a particular trait in an organism is referred to as

A. A genotype
B. **A phenotype (correct answer)**
C. An allele
D. A chromosome

MCQs query students about the rationale or causation of key terms in their curriculum. Inverse Definition MCQs have answer options that are succinct concepts (usually a noun phrase). Other MCQs require distractors that tend to be longer, whether they purport to be a list of attributes or explanations of how a process occurs. Inverse Definition questions present a straightforward way of describing a term, or question target, and then asking for that target.

In the QA, Question Generation, QG, and Educational Theory communities, the part of the question that presents the query is called the "stem." In Example 2, the stem is, "What is a phenotype?" Some people might consider the stem to be the actual question, but when the term "question" is referred to in this work, I am describing all of the information in the question, that is, both the question and its answer options.

Answer options include the correct answer (B in Example 2) and the distractors (A, C, and D). In instances where there is additional information available concerning the question, such as explanations of the answer, ratings of the question, and question topics, I call this information "related question materials."

Assessing IDQs presents a challenging computational linguistic task that is important to solve because of the high prevalence of these question types. These common IDQs have a complex structure and differ from factoid and list questions, which have been the primary focus of traditional QA systems [6].

Questions of the form "Who is Columbus?" or "What is a Tsunami?" that seek to provide definitions have been considered in Question Answering competitions funded by National Institute for Standards and Technology (NIST). In contrast, IDQs have not been included in such competitions including TREC. The correct answer to an IDQ needs to cover all components mentioned in the given definition, for example

Example 3

A *compound* [a] that is *synthesized by* [b] *both humans and geranium plants* [c] is known as

A. Cellulose [a]
B. **ATP [a + b + c] (correct answer)**
C. Ethyl alcohol [a + b]
D. Chlorophyll [a + b]
E. Mercury [a]

As shown in Example 3, there are three relevant components [a], [b], and [c] that a correct answer needs to contain. "ATP" is the correct answer because it satisfies the requirements of being a compound that is synthesized by both humans and geranium plants. The other answer options, known as "distractors," are missing at least one component of the correct answer term. For example, cellulose is a compound [a] but not synthesized by humans or geraniums (not [b] and not [c]). Even less relevant is mercury, which is indeed a compound, but also a "heavy silvery toxic univalent and bivalent metallic element" and thus very unlikely to correctly answer this question [9].

Of the set of five answer options, A through E, there are two weaker answer options, the compounds cellulose (A) and mercury (E). Would this be a more difficult question to answer if these answer options were replaced with terms closer in function to answer options C and D? The answer options are all related to the original question, but the closeness of the answer options to both the question and each other steers the difficulty

of this question. In order to correctly answer an IDQ, all of the components of the question need to be satisfied.

1.1.1 Quality and Differentiation with Multiple Choice Questions

The ideal MCQ is answered by a student who retained the information from the curriculum and is not easily answered by a student who did not. The question helps differentiate the good students from the bad, or gauge where a student is in his or her academic progress. From an instructor's viewpoint, questions need to be difficult, but not too difficult, because if a good student cannot correctly answer the question, there is a problem with the question. There are two general areas where the MCQ could be faulty. The first is in the question statement, or the stem of the question. Perhaps the stem is unclear, misleading, or simply lacks a distinct, well-scoped answer. The second way that a MCQ can be weak resides within the answer options.

If the question stem and answer options are appropriately aligned to the query topic, the result is a functional MCQ. A good MCQ is a question that combines two features: difficulty and discriminating power. Thus, a good question covers the topics presented in the curriculum at the correct level of difficulty for the course. A good question communicates to an instructor and the students themselves that the subject matter is understood.

In my work, the primary focus is on a distinct type of question that presents descriptive information about a topic and asks for the name of that topic. Questions of this type were filtered to include those that describe a biological concept or process. Filtering also controled the quality and consistency of the question statements or stems as the questions were processed. Armed with data that was filtered for quality and descriptive type, I look to address what makes a good multiple choice question distractor, or a good wrong answer.

Example 4

Which hormone secretion pattern is directly affected from jet lag?

A. **Cortisol (correct answer)**
B. Insulin
C. Thyroid Hormone
D. Adrenaline
E. Calcitonin

Example 4, has a difficulty of 56% which is considered a more difficult question in the question set used in this research [10]. The question was answered 282 times and 192 students chose the correct answer A. What makes this question difficult? Is it that all of the answer options are hormones? Is it that the term "jet lag" might be a paraphrase for a condition that this sought after hormone controls? If this question presents a desirable level of difficulty, what about it could be replicated in other questions to achieve the same desired effect? Modeling good question difficulty and the many ways that a question may be difficult will be discussed further in Chapters 3 and 4.

If a MCQ is not correctly answered by a majority of the good students in a class, the answer options must be further analyzed. The same logic holds when modeling the low-scoring students in a class. If a question is easily answered by all of the low-scoring students, it is not difficult enough. Modeling how a high-scoring student will answer questions in comparison to how a low-scoring student will answer them is discussed further in Section 3.3. The goal of creating a good MCQ is to include distractors that are effective at distracting many of the average or low-scoring students, but do not succeed in garnering more attention than the correct answer does from the high-scoring student. A difficult question is one where the correct answer and the distractors are closely clustered in terms of the ratio of high-scoring students to low-scoring students who chose them.

After analyzing what makes a good MCQ, the higher-level question becomes: "What makes a good exam?" Naturally, a good exam is filled with good MCQs, but more rigorously, the three components of a good exam are

- appropriate question difficulty
- sufficient question discrimination
- complete topic coverage

The difficulty of the questions must be gauged so that they accurately measure whether students comprehend the course curriculum. If an exam is secondary school-level, it should not be asking university graduate-level difficulty questions. Again, the goal of exam-taking is to show understanding of instruction so that a student may graduate to more difficult topics (which usually depend on comprehending aspects of the previous, understood instruction). Including questions that differentiate between the comprehension levels of students is another key component.

Finally, a good exam covers all of the topics in a course's curriculum [11]. In a discipline like biology, this means covering all of the processes, structures, functions, and key terms that were covered in the course or in the section covered by the exam.

This research focuses on optimizing exams to contain good questions that are both appropriately difficult and effectively discriminating. While vital to good exam building, complete topic coverage is outside of the scope of this research. A speculative answer to the "complete coverage" issue, and one I believe that is accepted by most professional testing organizations, is that a random sample of questions that cover all topics is sufficient to judge the quality of a test taker's knowledge.

1.1.2 The Automated Analysis of Multiple Choice Questions

There are two recent popular question answering programs that reveal important research foci. Apple's Siri system looked at open-ended questions while IBM's Watson answered quiz-style questions (without answer options). Both of these systems depend on a mixture of big-data machine-learning and context-driven, pattern-based methods to answer questions. While not perfect, they show the power and possibilities that large-scale question answering systems can provide. I am applying similar technology (albeit on a far-smaller computing scale) to a different type of question.

In this thesis I use data from the Web as possible answers to MCQs. In determining what is the correct answer to a question, I utilize a series of text comparisons that are also used by many question answering systems. Other systems, such as IBM's Watson, also use information from the Web. Many search tools incorporate latent semantic analysis (LSA)-based approaches to indexing and retrieving similar text [12]. One particularly relevant example is in the automated TOEFL challenge where many systems used LSA to help determine which answer options were correct [13]. As I noted earlier, as successful as the these approaches might be, they are unable to measure a particular question's difficulty or discriminating power. These approaches will be addressed in Chapter 2.

1.2 My Contributions and the Thesis Outline

The purpose of this research is to use current language processing techniques and online search tools to aid in discovering what differentiates a difficult-to-answer question from a more straightforward one. In this research, I use a type of complex question previously unexplored, the Inverse Definition Question.

MCQs used in measuring academic aptitude and subject comprehension are an extremely useful tool as they play a gatekeeper role in higher education admittance and

professional accreditation. Simply put, they are valuable commodities and as such they are protected intellectual property. The value of the proprietary questions authored by professional testing organizations is second only to the data associated with how students actually perform in these exams on a question-by-question basis. The educational testing groups also develop algorithms based on past student performance on questions that determine what questions are given next, on the fly, to test takers. This is especially true in computer-based testing environments.

The questions, student results data, and question-queuing methods employed by professional testing companies are not available for public research for fear that this proprietary information might be used by students to game their examinations in some way or by competitors in the testing world. In this thesis I replicate the professionally authored questions with ones developed on a crowdsourced question creation website. These questions are written and peer-reviewed by students. The new questions are then answered by students, and the results of their performance are available on a question-by-question basis [10]. In this work, the algorithm used to measure question difficulty and differentiation power is a standard in educational testing and while perhaps not the method used by the testing companies, its underlying logic produces useful results.

The anonymized data used in this research may be obtained for academic and open source research by contacting Paul Denny of PeerWise [10]. Similarly, the anonymized data is available from this author in csv format.

In this thesis I look at two distinct data sets to analyze question difficulty and discriminating power. The first is a set of 1000 questions authored by professional test makers, called the *test set*. The second is comprised of two crowdsourced exams from student-authored PeerWise data, called *Set1* and *Set2* respectively. Although this data only samples a subset of the total questions available in this field, it provides sufficient data to analyze question difficulty and discrimination power. Of course, further research could extend both the number of questions and the analytics proposed in this dissertation.

Though my experiments in exam building and developing a web search-based query pipeline, I show that useful exam evaluation data can be built from online, crowdsourced question banks. I also show that the automated system, using basic matching algorithms is more than twice as effective as random selection at finding the correct answer to a question.

I provide a general background to the research areas that impact question answering and Inverse Definition MCQ difficulty and discrimination in Chapter 2. In Chapter 3,

I describe the experimental design of an automated question answering and difficulty measuring system. I also discuss how this system extends earlier experiments. Chapter 3 describes automatically building exams based on sets of students who have answered questions in common. Chapter 4 presents question difficulty and discrimination results from both the automated system and human test takers. Finally, in Chapter 5, I give conclusions based on the work so far and present an overview of plans for future work.

There is also a series of additional support materials presented in the Appendices. Appendix A lists all of the example questions used to describe IDQs in this dissertation. Appendix B presents a background on Question Answering that augments Chapter 2. Appendix C describes methods for discovering definitions in text referenced in Chapter 2. Appendix D presents background material on using adjacency matrices to build exams and is referenced in Sections 3.4 and 4.4.

Additional support materials also include Appendix E showing the MySQL queries that revealed parameters of the exam data mentioned in Chapter 3. Appendix F reviews the details of the automated question analysis pipeline summarized in Chapter 3. Appendices G and H show the results of Item Analysis, discrimination power, and distractor usefullness. Finally, Appendix I shows the results of one of the Lucene experiments mentioned in Sections 3.2 and 4.2.

Chapter 2

Background and Literature Review

There are four research themes that this thesis builds upon. These four themes make up the sections of Chapter 2. The first is the importance from an educational and training viewpoint of the automatic solution and analysis of MCQs. The second is the use of latent semantic analysis and other text similarity measures to answer MCQs by finding similar terms in different locations. The third is the use of freely available resources to augment improvements in educational instruction and comprehension measurement. This includes the burgeoning developments in online education and the related support materials needed for massive open online courses. The fourth research theme delves into how best to automatically measure question difficulty and discrimination. Much of this research comes from the development of question answering and generation systems. These four themes are discussed in the following sections.

2.1 Why Automated Analysis of MCQs is Important for Teaching and Learning

MCQ exams were first used on a large scale to test potential United States Army servicemen's intelligence before World War I. In the last 30 years, MCQ exams have dominated standardized scholastic testing in the United States [8]. Currently, multiple choice is a large part of the following exams [14] used both in the United States and internationally:

> The American College Test, American Services Vocational Aptitude Battery, Fundamentals of Engineering Exam, Graduate Record Exam for graduate study, Law School Admission Test, Medical College Admission Test,

> Multi-state Bar Examination, Scholastic Aptitude Test (SAT), Test of English for International Communication, and United States Medical Licensing Exam.

The MCQ corpora used in this research are from the New York States Regents Exam, the Advanced Placement exams, the College Level Examination Program (CLEP), the SAT, and PeerWise, an open source question bank. These tests are available in subject-specific areas such as biology and psychology (the subjects used for building the test corpus as described in Chapter 3) and have been given to millions of students [15].

MCQ-based exams are the predominant manner of testing used in the United States for secondary school exit exams and college and graduate school entrance exams. There are many advantages of well-developed MCQ exams over other forms of assessment. These advantages include efficient and unbiased grading due to the answer choices being clearly correct or incorrect. In addition, the person grading the exam need not be an expert in the field being assessed. A good MCQ exam can cover a great deal of course materials in a shorter exam than other testing alternates, such as short-answer written exams. Comprehensive coverage of large subject areas, lack of bias, testing and grading efficiency, and the ability to give these exams effectively in many environments, including online, has led to this format's dominance in standardized testing [16].

Within the area of education evaluation there are three research areas to which the automated analysis of MCQs is linked. These are

- increasing use of online educational resources and tools to best deliver high-quality instruction to large numbers of students
- value of student self-questioning, through exam-taking as an educational tool
- design of MCQs that are both high quality and highly differentiating.

One important academic trend in post-secondary education is support of in-class instruction with online materials. This spectrum of online support spans from basic web pages and email listserves to completely online instruction by highly regarded institutions [17]. While full credit online courses are still less common, many online tools allow many thousands of students to participate in a shared academic experience through automation of registration, remote lecture viewing, online bulletin

boards, and computer-based examinations. In addition to the multiple products offering high-quality instruction, now meaningful performance reviews are available from automated grading systems [18].

As the online education market attracts the attention of inventors and software developers seeking their slice of the automation pie, some software systems aim to support the learning environment in ever more innovative ways. Online companies such as Quora focus on aspects of instruction such as question asking and answering, but developed for wider and more sophisticated users than students alone [19]. One academic-focused question authoring and answering environment is Piazza [20]. Another such tool is PeerWise, "a web-based system that supports the creation of student-generated test banks of multiple choice questions" [21]. I use PeerWise data for the research experiments presented in this thesis because it is open source software. PeerWise is discussed in more depth in Section 2.3.

Coursera, a spin-out from Stanford University is the largest MOOC (massive online open courses) provider currently, followed closely by edX and Open Yale which are collaborations of Harvard and MIT, and Yale, respectively [22] [23] [24]. Coursera is focused on collecting "data to understand the learning outcomes from facilitated discussion," a facet of remote learning that is trying to be bridged through online support [25]. Andrew Ng, one of the creators of Coursera, along with Daphne Koller, run a course on machine learning that has moved from 350 students taking the course in a conventional classroom to hundreds of thousands via the internet [22]. The crucial question then becomes how to best leverage known approaches to measuring student comprehension when the scale radically increases.

The conventional model for creating an MCQ exam is for an instructor to study lesson plans and related textbooks to create the items and answer options used in a question. Questions are then given in an exam and after an instructor grades the results, he or she goes back to review how effective each question was in testing the student's understanding of the subject matter [26]. In other words, each new set of questions can only be tested by subjecting it to a test population and inspecting the results. Only then can an exam be deployed in earnest. This is a time and resource-expensive model.

Recent work has shown that the act of authoring questions corresponds "to higher order levels of cognitive skills in the Bloom taxonomy of educational objectives" [21]. In other words, students who participate in the process of creating questions in a discipline gain a deeper understanding of the question topic. "Findings indicated that even controlling for the students' prior knowledge or abilities, those who were highly en-

gaged in on-line question-posing and peer-assessment activity received higher scores on their final examination" compared to their peers [27]. "The results provide evidence that web-based activities can serve as both learning and assessment enhancers in higher education by promoting active learning, constructive criticism, and knowledge sharing" [27].

Engaging students with technology, whether their courses are online or not has "focussed on the growing conviction that students do better if they can discuss course materials" [25]. Traditional teaching methods are being "enhanced by web-based technology" and the grade-increase results for the students who participate the most is statistically significant [28]. Participatory learning environments such as PeerWise show that students who contributed the most by authoring questions, answering questions, writing comments, and submitting ratings had the largest improvement in their course grades [28]. The PeerWise-based research was performed on college-level biochemistry courses and another digital education project, an interactive, question-answering "intelligent" textbook based on a biology textbook [29]. Efforts to increase educational outcomes in science education are utilizing cutting-edge educational research. Additional research using PeerWise data includes incorporating gamification techniques to increase the stickiness of the online environment and increase student participation via small awards or "badges" [30].

Students authoring questions, evaluating questions produced by their peers, and explaining or commenting on questions are behaviors documented in an instructional methodology called Participatory Learning Approach (PLA) [31]. PLA is an extension of educational evaluation methods that focus on improving performance by using teaching tools in new ways supported by web-based technology. For example, recent research has shown that students, especially those from disadvantaged backgrounds, benefit from frequent testing [32]. This study also found that digital personalized quizzes "act as an aid to teaching" and reflected on how best to use their findings "as a large-scale prototype for how such testing effects can be exploited in the digital era" [32]. The reasoning behind performance improvements, according to the researchers, is that forcing students to test more frequently is akin to making students write down concept maps of what they remember from their curriculum.

Concept mapping has explored the relationship of student performance to studying behaviors or specific study tasks. In psychology, concept mapping is used to represent a given domain's knowledge or all of the concepts and ideas that make up that subject area. In a concept map, domain knowledge takes the form of understanding the main

ideas in a subject area and how they relate to one another. An expert in a particular domain will have a more abstract understanding of the interconnectedness of these concepts.

Thus, if a teacher is considered an expert and a student a novice, one way to examine students is to have them construct visual representations, usually graphs, that link the main ideas within the subject on which they are being tested. The student's map is then compared, using a map-traversal algorithm, to that of the instructor's (or a gold standard). The Pathfinder algorithm, for example, is one of several algorithms used for traversing the edges or links between the concept nodes to empirically measure the distance between the maps [33]. This approach works particularly well in the sciences where there is a distinct curriculum of ideas and a specific set of terms that describe the concepts, functions, and processes in a domain. As Merchant notes, concept mapping addresses "one of the main problems involved in the teaching and learning of physics": the formulation of concepts [34]. Using concept mapping to test college students has been implemented and analyzed [33] [35].

Concept mapping requires an explicit and complete representation of domain knowledge which in practive is expensive and often prohibitive to construct. I discuss this approach and it's limitations in Section 2.2.1.

Measuring the effectiveness of tests and other educational metrics has been explored primarily in educational psychology, later in mathematics and statistics, and more recently in language technology applications. The evaluation of item difficulty via Item Analysis is common in Educational Testing Theory [36]. Item Analysis is commonly used by instructors who author their own questions. However, there are other types of testing measures that solve other specific comprehension issues.

Item Analysis itself is commonly called "Classical Test Theory" because the more data-intensive psychometrics of Item Response Theory (IRT) have proven useful in computerized adaptive testing (CAT) [37]. CAT is the model of testing where as a student progresses through an exam, the questions they are given are based on a decision-tree of how well they have performed on the most recent question. Current online standardized tests such as the Graduate Record Examination (GRE) and the Graduate Management Admission Test (GMAT) use IRT, which utilizes past models of student behavior to "give the probability that a person with a given ability level will answer correctly" [37].

Using Markov Chain Monte Carlo (MCMC) methods for estimating missing data in IRT has gained traction as a viable approach to dealing with testing situations that

have incomplete test data [38] [39]. This method for dealing with sparse data will be discussed in more detail in Chapter 3.

While Item Analysis and Item Response Theory are the underlying structures for most comprehension measures, there is ongoing research specific to using smaller scale online testing or tutoring systems than the GRE and GMAT. Computer based training (CBT) has provided the opportunity for testing of the Q-matrix method, which "mines student behavior to create concept models of the materials being taught" [40]. This approach aims to create effective feedback loops for remedial education where students fail to grasp some concepts in examinations. The Q-matrix method combines item analysis with another feature that maps each question to a concept topic and uses hill-climbing algorithms to automatically create the relationship between questions and concepts [40].

Avoiding placing too much weight on lucky guessing in multiple choice exams is the focus of [41]. Hensler uses the Cloze question format designed for question generation systems in MCQs [42]. While most research looks at latency, item selection, and response times in answering, my research considers *a priori* models of question difficulty and discrimination. The approach I implement in this research will be described in Chapter 3 following a continued review of other related research in this Chapter.

2.2 Research Related to the Automated Information Extraction from Text

The task of extracting "knowledge" from text has long been a goal of the AI community [43]: Two major approaches to this task have included using explicit knowledge representation techniques and employing some similarity measure. I have adopted a form of the second technology but before describing my approach I give a brief summary of the explicit knowledge representation including semantic networks, frame scripts, conceptual graphs, and FrameNet.

2.2.1 Explicit Knowledge Representation for Information Extraction

The underlying tension between knowledge representation formats and language-data oriented algorithms is that the former is used as a way to represent how the world

should be (organized by formalized logical relationships) and the latter at how the world is (characterized by actual occurrences of text). Knowledge representation depends on some aspect of overtly creating relationships that can be queried while a language similiarity approach uses the patterns and proximity of words based on actual occurrences. A limitation of the word similarity approach is that it has to use approximations to deal with previously unseen word occurrences which will always exist. The evolving nature of language, if nothing else, makes word similarity approaches less perfect than they might initially appear.

Historically, explicit methods for representing concepts and relationships have been shown with graphs where nodes represent the facts and the arcs represent the associations between these facts or concepts. A *semantic network* is a term used to describe a family of graph-based representations. In [44], each node was a word concept and each meaning of a word was represented as a graph. An ambiquous word, could have several different graphs. Then, "scripts" turned the graphical relationships into a semantic language that could be used to tell simple stories about birthday parties or going to a restaurant based on the roles, conditions, track, props, and results [45]. Implementing these scripts was dependent on matching key words in the text.

Natural language understanding has benefited from early semantic networks, some of which were modeled on dictionaries and defined words in terms of other words. In the case of conceptual graphs, John Sowa of IBM created a predicate calculus-like language that was intended to compute (again *a priori*) relationships in the world [46]. At the same time, researchers at University of California, Berkeley created an explicit framework of concepts and relationships called Frame Semantics [47]. An extension of this approach is the ongoing FrameNet project at Berkeley. FrameNet attempts to capture the semantics of actual world situations in a manner that is both human- and computer-readable. The goal is to build a lexical database of English by annotating examples of how words are used and relate to one another in actual texts. The relationships are "based on a semantic frame: a description of a type of event, relation, or entity and the participants in it" [48]. FrameNet remains a useful tool in semantic role labeling.

Another semantic network, concept maps, have also been used in question generation [11] and is a type of abstract relationship management tool that is related to the ontologies used in Wikipedia tables of contents and topic sections that are utilized in some Question Answering and Question Generation systems. One motivation for using concept maps is the hypothesis by some psychology theorists that in humans

"questions are generated from a knowledge representation modeled as a concept map" [11]. Concept maps aid in developing questions that fully cover the curriculum of a class. Some online educational environments use "scaffolding" techniques that seed question topics into question sets, a PeerWise-based comparison of the topics organically contributed by students to those deemed essential by instructors had complete overlap [49]. Curriculum coverage is an important part of developing good exams, but it is outside the focus of this research project.

WordNet is a hybrid semantic label database and a cognitive synonym (synsets) relationship network consulted for discovering the semantic closeness of concepts [50]. WordNet relations combined with corpus statistics are used in many natural language processing systems including those seeking paraphrases and snyonyms of terms to expand potential matches when comparing words that differ in their surface representation [51]. In Mitkov's work generating MCQ distractors, WordNet is used to discover distractors that were semantically close to a question's correct answer. The distractors produced automatically were found to be better than those created manually [52]. Finally, I also use WordNet weights in the matching algorithm of the automated question answering and difficulty system to increase the likelihood of paraphrase and synonym matches.

2.2.2 Extracting Knowledge Through Similarity

The concept of similarity is straightforward: "How close are two or more things?" This question is usually clear if the things being compared are numerical in nature. They are less clear when the comparison is based on bit strings, graphs, or words. Many researchers in language technology use several similarity metrics in an attempt to be exhaustive in their comparisons [53].

A problem with similarity measures is that there are dozens of algorithms for measuring similarity, but many are domain-dependent or depend on sophisticated assumptions that are difficult to compare when a researcher is seeking a more universal similarity metric [53]. When closeness is being compared in instances that are not numerical in nature, there are assumptions that can be leveraged to model measurable closeness. From a data-driven language technology perspective, closeness requires using mathematical models of language in order to produce a result from a similarity comparison. An information-theoretic approaches helps to decipher the complexities [53].

The comparisons used in this research are based on the "bag-of-words" (BOW)

or latent semantic analysis (LSA) models. Bag-of-words models take strings of text, often stem the words, and then use them for comparisons based on all of the terms in the bag, often including term duplicates and usually excluding word order. This is a simple way of looking at matching terms without regard to sequence, (which of course, conveys a great deal of meaning). Bag-of-words models are sometimes the first step used when comparing terms and in this research, a bag-of-words based approach was the basis for the Rouge software which is used as a benchmarking system described in Chapter 3 with the results in Chapter 4.

There are three broad types of vector space models (VSMs): term-document, word-context, and pair-pattern matrices [54]. The main concept behind a "VSM is to represent each document in a collection as a point in space (as a vector in a vector space). Points that are closer together are assumed to be semantically similar and those that are far apart are semantically distant. The user's query is represented as a point in the same space as the documents (the query is a *pseudo-document*). The documents are sorted in order of increasing distance (decreasing semantic similarity) from the query and then presented to the user" [54]. A more modern version of the BOW approach is Lucene. The Lucene software is an extension of the BOW method used by the Rouge program run on my "test set" and includes matching scores that are weighted based on the lengths of the strings [55] [56].

I employ LSA-based term-document vectors to compare the terms in a question stem to those in the returned definitions from the Web for each answer option. The terms in the question stem are represented as rows in a sparse matrix where the columns represent the document made up of all of the terms in the definitions from the Web. The implementation of this approach may be found in Section 3.2.2 and the results in 4.2.1.

LSA assumes that words that are semantically close will occur in similar text [12]. LSA was famously tested as a method for solving multiple choice TOEFL synonym exams where it was used to successfully answer 64.38% of the questions [13]. The TOEFL synonym MCQ set was originally produced to meausre English language aptitude of non-native speakers of English. Thus there are average human performance results for this TOEFL MCQ set and researchers used this question set as the basis for a research challenge. There are 80, 4-option MCQ in the exam [57]. The average non-English US college applicant scores slightly higher than the LSA method, answering 64.50% correctly [58]. The TOEFL test set was answered 100% correctly in 2012 using Principal Component vectors with a Caron P algorithm after years of

incremental improvements using a selection of algorithms [58]. The TOEFL question set is a research standard that encouraged the use of a selection of algorithms on this task. The IDMCQs sets used in my research require more sophisticated concept-based knowledge to be answered than leveraging synsets.

Several times in the previous sections I have referred to the term "similarity" when describing the relationship that I am seeking between the question statement and the definitions associated with the answer options. By similarity, I mean lexical semantics, or the relationship that the underlying meaning the components of words and lexemes have to one another. For the purposes of this work, it refers to the closeness in meaning that one string of text has to another. Through a mixture of many lexical characteristics such as synonymy, paraphrase, and metaphor, two pieces of text may have very close meaning, but reflect very different surface representations.

Lexical similarity measures are important to this work because I compare text strings to one another when testing the similarity of the answer options to the original question stem. This similarity testing is described in more detail in Section 3.2.2 and the results are presented in Section 4.2.1. I am seeking to map the definition of a concept presented in a question to that of the definition of the answer options. Ideally, the answer option that correctly answers the question would contain the exact same text as the question. But since that is an unusual occurrence in language, being able to map word variants to one another is very useful in determining whether two texts are really talking about the same thing.

There are many ways to introduce additional information about words into text comparisons. One is to incorporate WordNet lookups with content-based statistical models of word occurrence. WordNet is a database of lexical relations that depends on distributional semantics to link related words [59] [51], and is described in Section 2.2.1.

BioWordNet, which initially appeared to be a relevant network to incorporate, was not used. After using WordNet and attempting to implement BioWordNet, and considering other research that unsuccessfully attempted to include BioWordNet [60], I decided the tool was unsuitable.

Definitions, including those for the types of biological terms described in the MCQ data in this research, follow a series of sentence construction patterns when they occur in text. Information Extraction components of QA systems incorporated algorithms that delineated a series of lexical patterns that are often used when an author is presenting definitional material in text. These patterns include appositives and copula

constructions, propositions, relations, and structured patterns such as a rule devised by Xu:

> "<TERM>,?(is|was)? Also? <RB>? called|named|known+as <NP>" [61].

In the instance of identifying definitions, the same lexical clues that are found in the MCQs are also found in the definitions on the Web. Search engines use the construction patterns, among other information, in their definition retrieval shortcuts which help users only find definitions when employed. Shortcuts help supplant the analog use of dictionaries or encyclopedias and have been effective in streamlining the delivery of a constrained type of result. One such shortcut is "Define: X" where "X" is the term being defined. The implementation of the "Define: X" approach is presented in Chapter 3 and a further discussion on definitional patterns is found in Appendix C.

Generating MCQ exams contains a number of steps, but the two major components are generating question topics from a domain and generating relevant answer options from that same domain. Generating MCQ test items was the focus of Ha's paper [62], describing a demonstration system she built while at Wolverhampton. Her system uses linguistic patterns to extract terms that name major topics from a document describing an academic subject area. Then, she uses a Wikipedia-based interface to aid in the human generation of distractors as the users navigate through a topic-based ontology GUI [62]. Similar, statistical approaches have also been implemented, with Foster running "an extremely small experiment" that augments controlled language topics with those automatically deemed significant from processing study guides in the assessment domain [63].

> "Similarity patterns employed in the procedure of selection of distractors are collocation patterns, four different methods of WordNet semantic similarity (extended gloss overlap measure, Leacock and Chodorow's [59], and Jiang and Conrath's [51], as well as Lin's measures), distributional similarity, phonetic similarity as well as a mixed strategy combining the aforementioned measures. The evaluation results show that the method's based on Lin's measure and the mixed strategy outperform the rest, albeit not in a statistically significant fashion" [64].

Mitkov et al.'s work, [64] [52] [65], was based on 144 "items" (answer options) being generated for a total of 20 questions based on the domain of a university-level course in English linguistics. The resultant test was taken by 243 United Kingdom and European university students and to evaluate the quality of the exam items, the researchers used standard "item analysis." Item analysis is derived from "classical test

theory" and provides information as to how well each item has functioned in the test [64] [36]. In this case, item analysis consisted of:

- the difficulty of the item
- the discriminating power
- the usefulness (or effectiveness) of each distractor

An evaluation method called "item difficulty" which is based on item analysis will be discussed in more detail in Open Source Question Answering Data, Section 2.3, and in Chapter 3.

Computational linguistic research into MCQs has also included efforts to automatically classify the difficulty of individual MCQs as described in Barker's work [66]. Other efforts include answering MCQs with a purpose-built multi-lingual QA module. Awadallah et al.'s study [67] showed the superior merits of using returned web snippets over web hits to answer questions using key term search, an approach that is a part of the pipeline described in Section 3.2.1 [67]. The question types they examined, which were from the "Who Wants to Be a Millionaire?" television quiz show, have a different form from the IDQs. This work examined multi-lingual MCQs and focused on factoid questions. Lifchitz et al. [68] sought to test the use of Latent Semantic Analysis (LSA) in French-language biology question answering. The questions were based on a seventh and eighth grade curriculum and their system had results similar to those of the students. In addition, they developed an original entropy global weighting model of the answer options for each question to return results equal to those of the students [68].

2.3 Open Source Question Answering Data

The use of standardized comprehension or aptitude exams in research requires having access to sets of exam data, which include the questions and detailed, question-by-question results from thousands of students. Unfortunately, such ideal data is very difficult, if not impossible, to obtain. The use of crowdsourced, human-annotated, or "human-in-the-loop" data has emerged as an important resource for human judgments including answering exam questions [69]. For example, Amazon's Mechanical Turk [70] and the crowdsourcing company Crowdflower [71] both provide avenues to gather human judgments on myriad tasks [72]. More specific to my research task,

there are other question authoring and answering environments available, which include question banks of test questions, including Piazza [20] and PeerWise. I have chosen PeerWise for this work; PeerWise has a relationship with the University of Edinburgh supporting several courses.

Crowdsourcing presents an alternate method from academic, institutional, and research-oriented document annotations for gathering useful, human judgment data. Annotations are human judgments on text and include comments (i.e., multilingual translations, part-of-speech tagging, and intentionality clarifications) that add supplemental knowledge to text and aids in the building of machine learning-based algorithms. Annotating words, phrases, and documents is the basis for many algorithms that support research in Computational Linguistics (CL) and Natural Language Processing (NLP), but the annotation process is expensive, time-consuming, and often purpose-built for just one task.

Extensible data sets that rely on microtask-built data transform the way human judgment data is incorporated into problems facing areas as disparate as educational testing, disaster remediation, and marketing surveys. While the applications are myriad, the techniques are new, and somewhat opaque. I present an example in educational testing and discuss possible applications in other domains in Chapter 5.

The supportive findings in student question authoring motivated the creator of PeerWise to build the question authoring and answering environment, or "test bank." Analysis of how students' use the PeerWise system showed that "it is the contribution aspects of the system, rather than drill and practice" that resulted in improved grades by participants [21]. PeerWise creator Paul Denny examined data from the system and showed that the act of creating questions, and specifically participating in discussions concerning the questions and their topics, was more valuable to a student than solely answering questions. Thus, a task that is so time-consuming and abstract for instructors is similarly taxing for students and forces them to understand topics for which they are constructing questions.

While the supportive literature reflects student questioning experiments conducted in computing classes [21], in graduate-level MBA courses [27], in literature [73], and in physics [34], one advantage of the PeerWise system is that it is a platform for the type of question-posing and answering that is a part of any academic discipline. Denny's paper focused on motivating PeerWise to provide a meaningful impact on student performance. He tested the utility by having four university-level computing classes use PeerWise as a study aid. The students who participated most avidly in the

system by authoring questions and answering questions performed better on the exams. Most important, "average comment length appears. . . to be significantly correlated with exam performance" [21].

The quality of the questions authored by the students is neither equal to those composed by the instructors nor commensurate with the quality of questions expected in university-level courses. This is logical since the students are novices at writing questions and their teachers have a great deal of experience. That being said, the quality of the questions is very high, especially so in questions that have been refined and improved by incorporating comments and suggestions from classmates, as PeerWise facilitates. The comment screen can be seen in Figure 2.5 for further discussion.

Denny is also focused on making PeerWise as fun and engaging an environment as possible to produce self-perpetuating student involvement in the question banks. Current work on improving the site is taking inspiration from gamification and meeting small challenges that are rewarded with badges. Gamification is seen as a way to introduce "stickiness," or the addictive component of some games into otherwise mundane activities such as studying for exams [74].

The PeerWise environment allows for students to tag questions with the topics that the question covers. Students may add new tags after they compose a question, or they can choose from the existing tags. Instructors, on occasion, seed the tag section of a class's question topic tags to encourage a broader coverage of questions to be authored. This is shown in Figure 2.2.

The PeerWise web interface is a logical extension, if not a literal child, of long-standing question answering environments such as list serves and community electronic bulletin boards where questions are posted by "newbies," or new members, and answered by longer-standing members, or "experts." The ease of use, well-designed graphical user experience and focus on creating MCQs produce data that is useful for question discrimination and difficulty research. Further, this type of tool scales well for student-self education and community building in the recent online education boom that aims to educate vast numbers of students remotely [17].

Developing a collaborative exam building environment, such as PeerWise, has resulted in more than just a set of potential exam questions associated with the related curriculum. The students self-police the quality and correctness of questions via a rating system and comments section. Question difficulty is ranked from 1 to 3, 3 being the most difficult, and question quality is measured from 0 to 5, 5 being the highest. The notion of question quality or "goodness" is more subtle than difficulty judgments.

It may pertain to how useful the question was to the examinee. Perhaps there is an intuitive sense of a well-formed question that is akin to other human perception tests. These question ratings are saved and can be used as a comparison to how the students actually perform on the questions. Further, creating questions forces students to understand concepts adequately enough not only to make a correct statement, but also to find good distractors that indeed distract their classmates. In classes that use PeerWise, instructors may make creating questions a voluntary, mandatory, or graded component of their classes.

In PeerWise, there are three steps for authoring and answering questions. The first is to write the question using the provided template. The template provides a window for typing the query and input cells for up to 5 distractors, A through E, as seen in Figure 2.1. Then there is space for a student to add an explanation of the question and refer to the related curriculum concepts, textbook section, or course notes via tick boxes.

As shown in Figure 2.2, these suggestions are based on previous links made between other questions and the course material. Topics may also be seeded by instructors to encourage questions that cover the entire curriculum; (note these topic tags are not hierarchical). There is additional input space for describing why the correct answer is the best answer choice. This is especially useful in cases where the answer options are closely related topics. Students may contribute multiple questions and all questions that they create are linked via a unique but publicly anonymous id.

The second component of the web tool allows the students to have a test-like experience by answering questions, as seen in Figure 2.3. The answer screen looks like that of a conventional online test, with tick boxes associated with each possible answer. An example of the PeerWise biology data is:

Example 5

What is the name of the areas between osteons?

A. Canaliculi
B. Lacunae
C. Lamellae
D. **Interstitial lamellae (correct answer)**
E. Volkmann's canals

When a student is presented with this question, a statement at the top of the page notes how many people have previously answered the question (244 other classmates

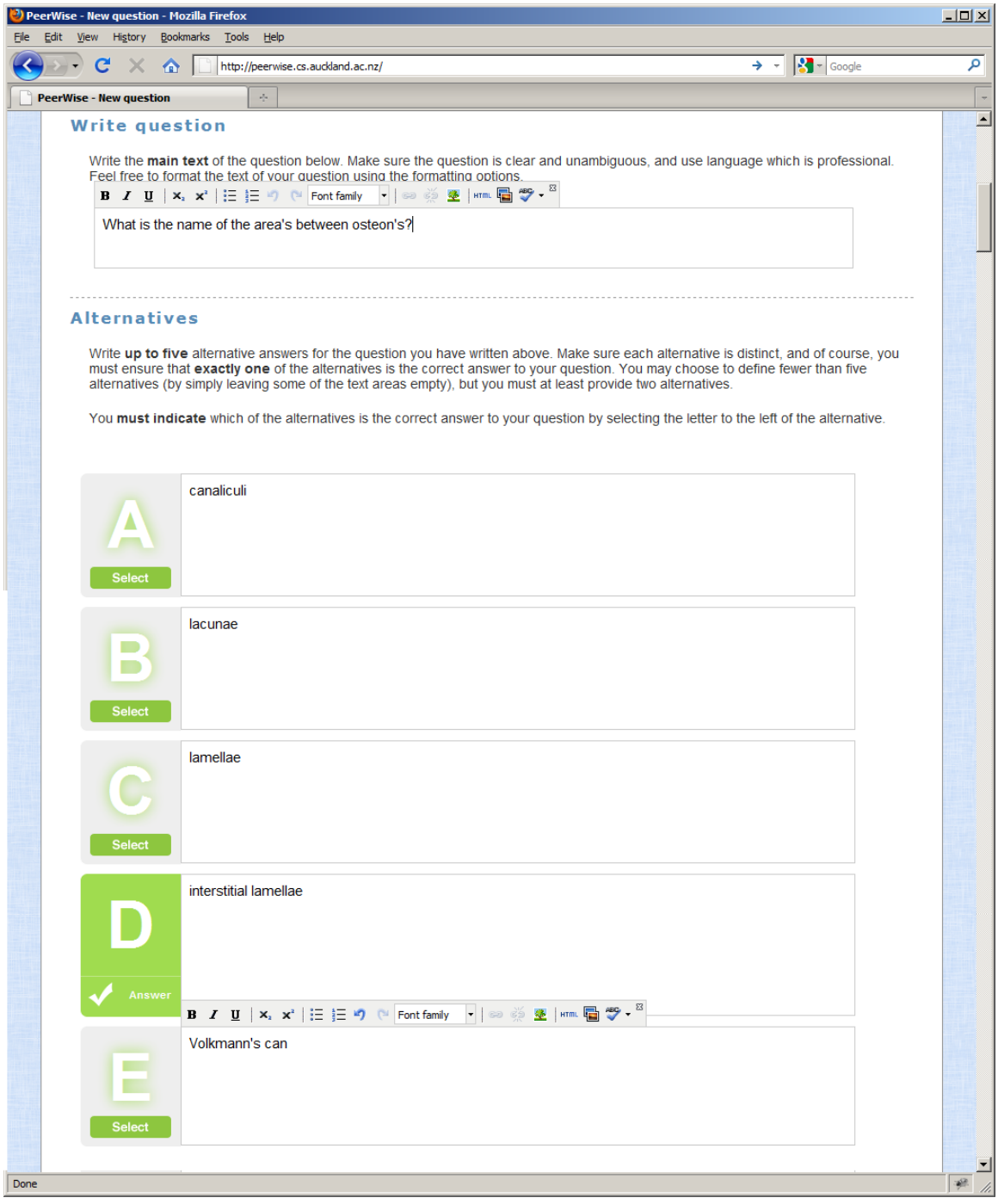

Figure 2.1: The first step in creating a question is to add the question stem text into the *Write question* window. In the *Alternatives* section, a student adds the answer options and highlights what he or she considers to be the correct answer.

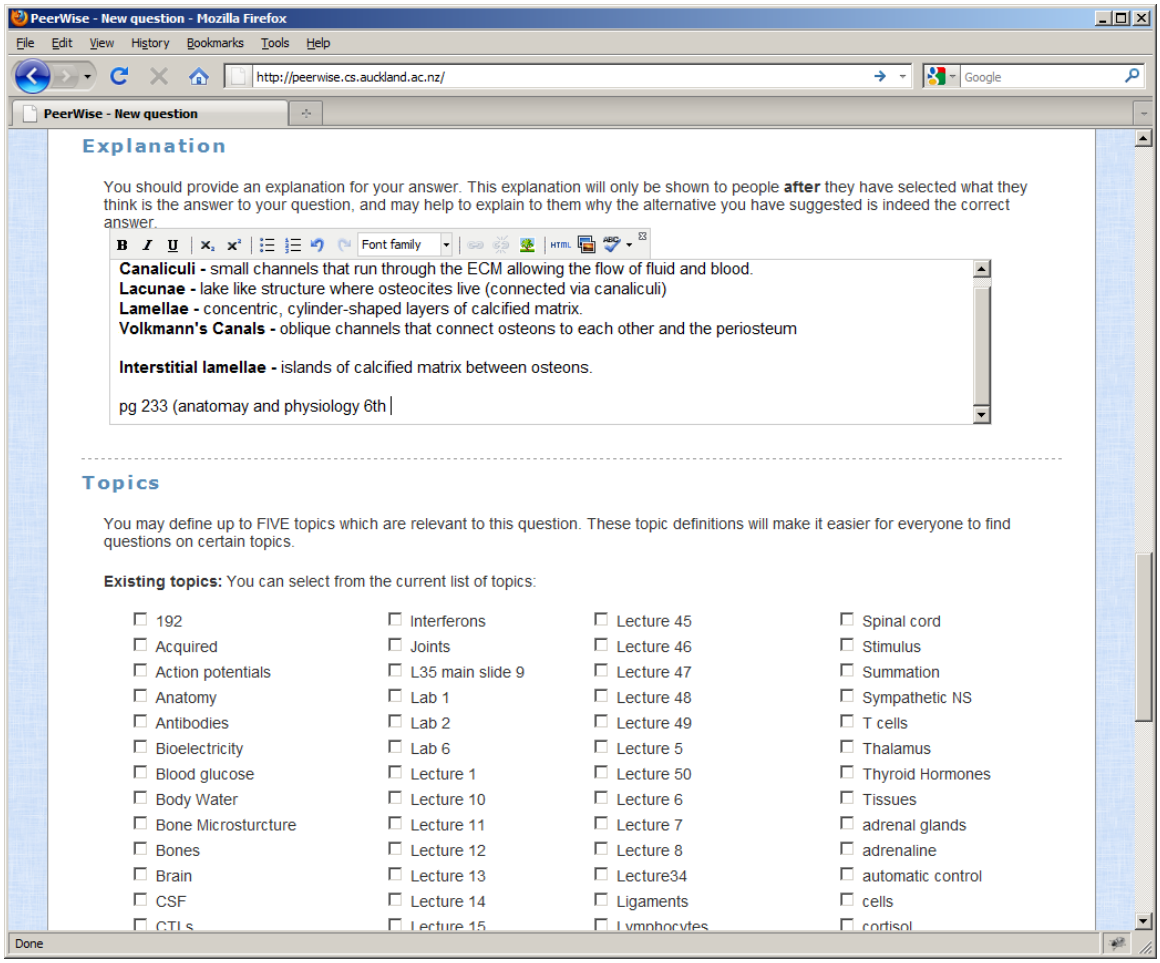

Figure 2.2: The *Explanation* window provides a text box for motivating the correct answer and often includes descriptions of the other answer options. Then, lower on the page are a number of tick boxes with listed options of topics to associate with the question and to be chosen by the student from his or her perspective.

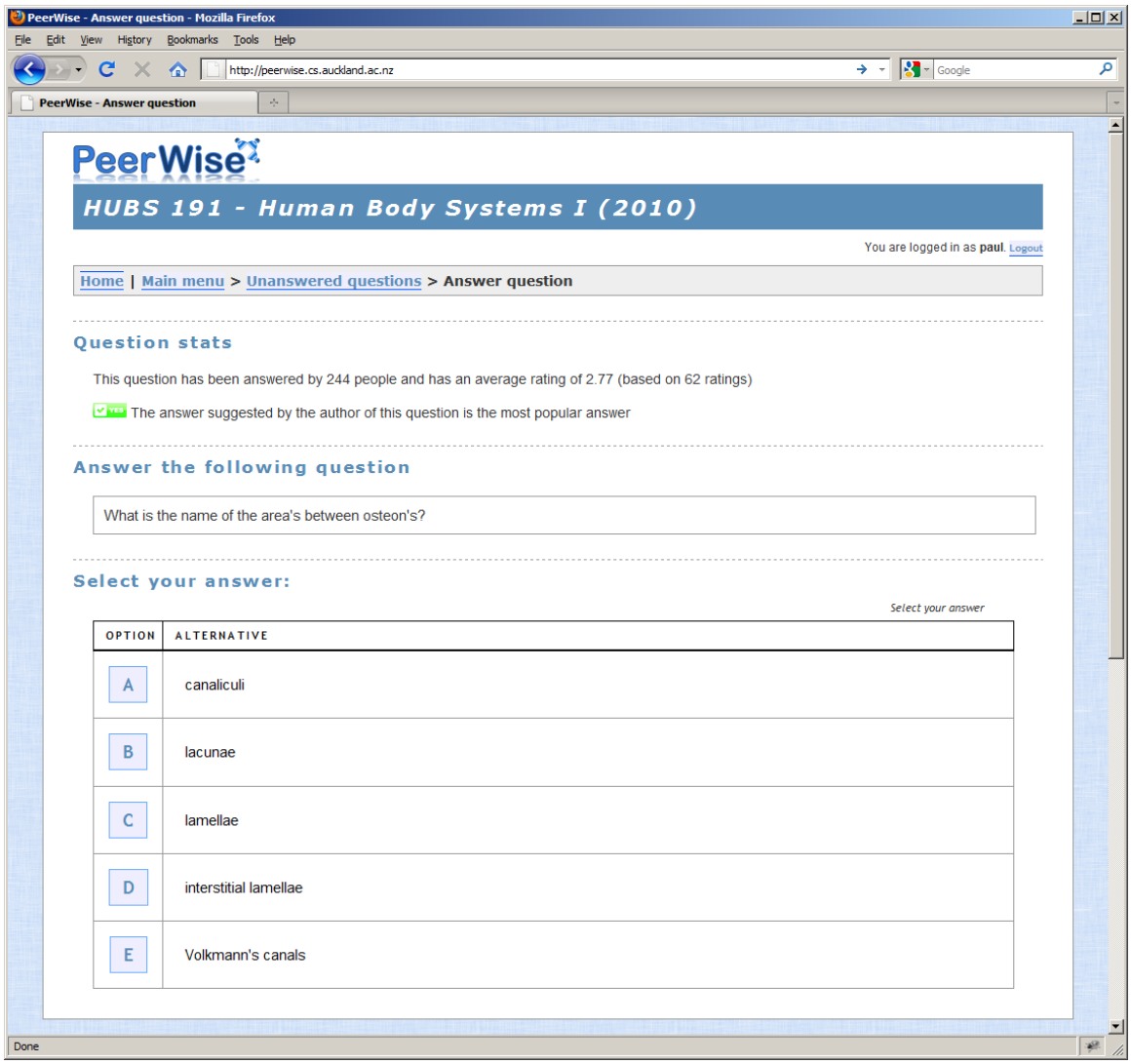

Figure 2.3: Here is an example from the PeerWise site of a student answering a question. The question-taking page notes how many students have currently answered the question and how high in quality they rated it.

in this case) and the average quality rating they gave the question (2.77). Not all of the students rate all of the questions; in this instance, the ratings are based on 62 responses.

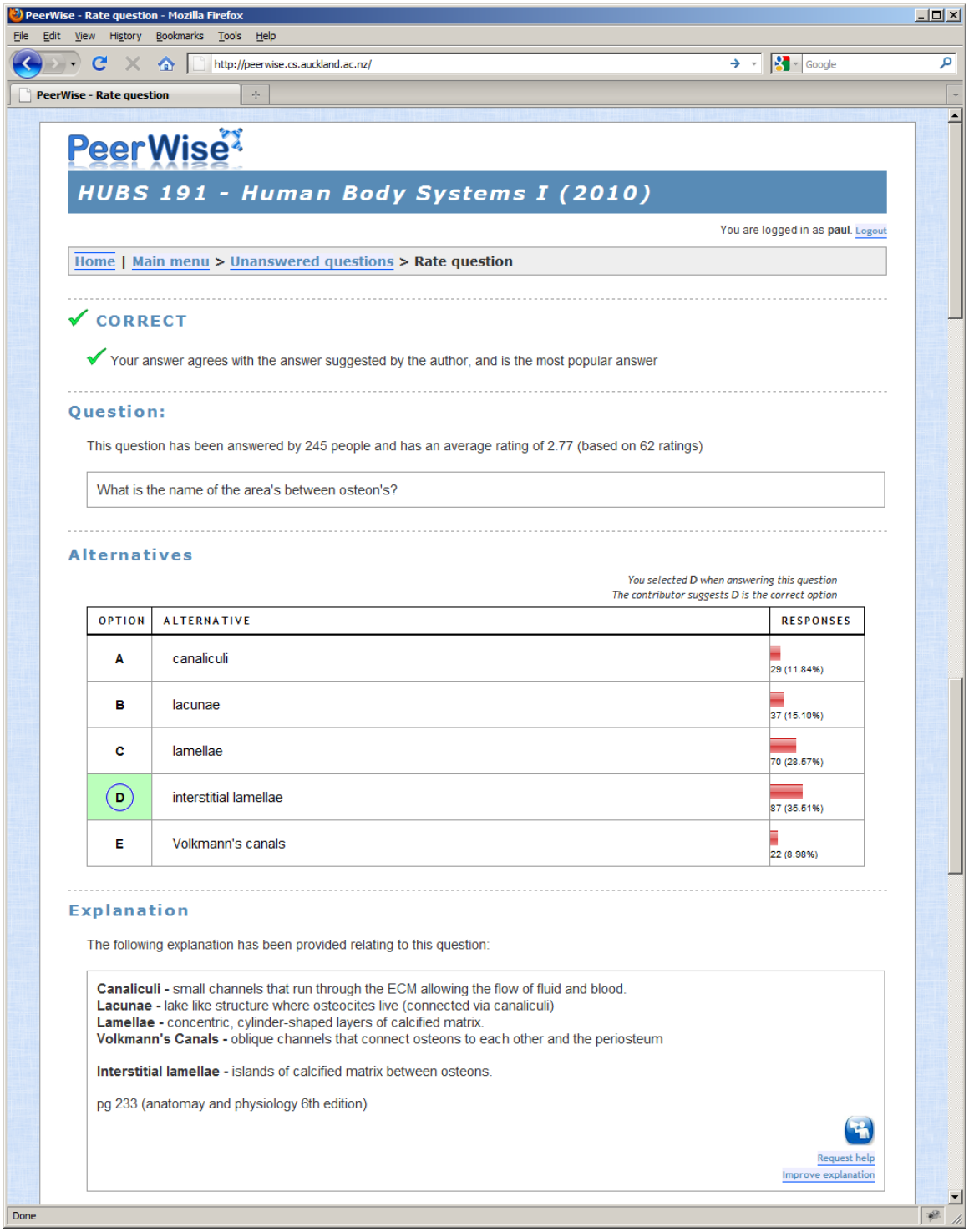

Figure 2.4: Screen shot from the student-examinee's perspective after choosing a correct answer. Once the student has answered the question, there is an immediate indication of how well the student did in comparison to his or her peers.

The third step is for the test taker to compare his or her answer choice to the correct one and then rate the question (as presented in Figure 2.5). For Example 5, the correct answer, D, was chosen 87 times, or by 35% of the students, as shown in Figure 2.4. The number of students who chose each answer option is listed as well as the question explanation. Students are then given the chance to rate the question both in terms of overall quality and in regard to difficulty.

Finally, comments, suggestions, and edits may be included and these are emailed to the student who authored the question so that any flagged errors may be corrected. The format of the comment section is similar to that of an electronic bulletin board, so it allows classmates to discuss the questions. The question-rating interface is presented in Figure 2.5.

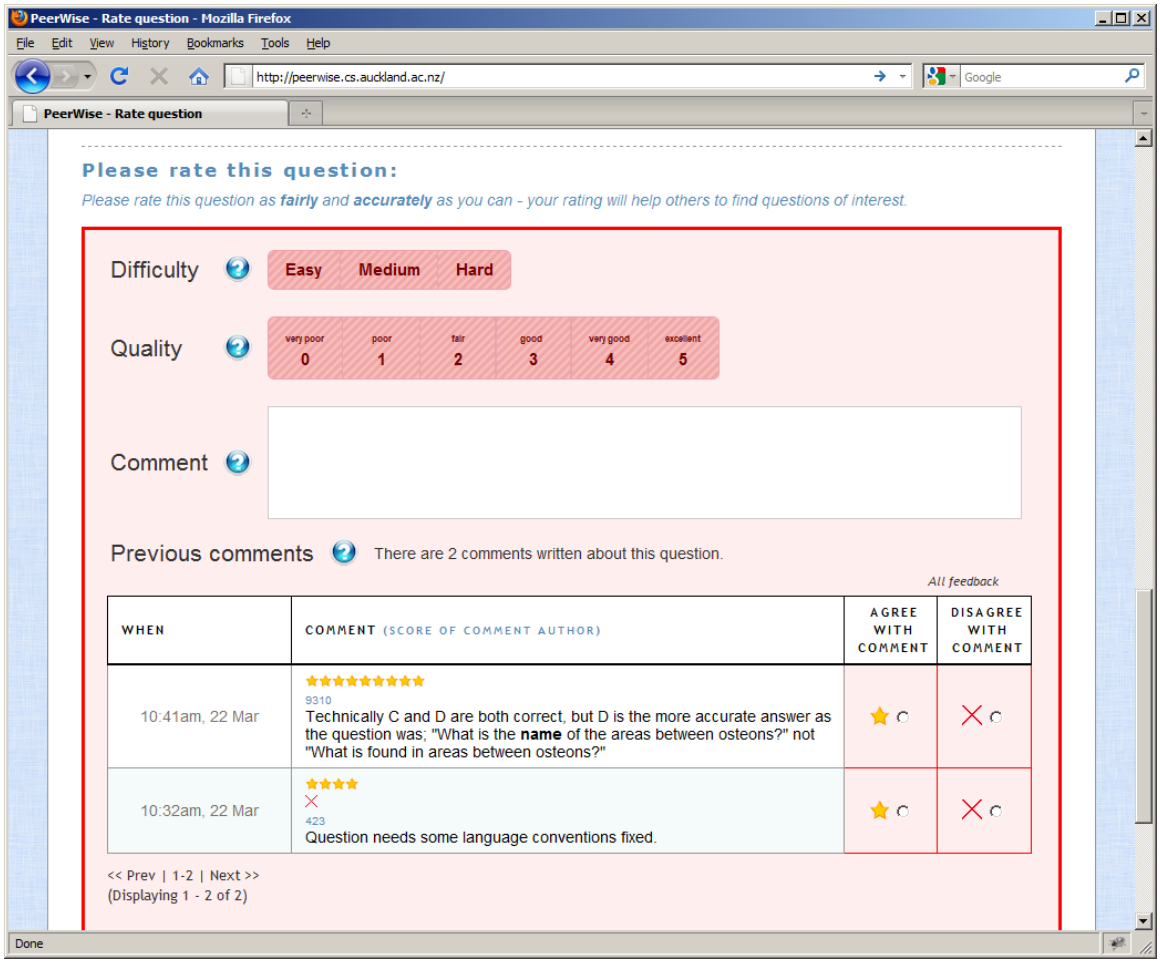

Figure 2.5: Screen shot of the question rating page from the student-examinee's perspective. Below the question rating is a list of comments that students have left when reviewing the question.

The PeerWise system was not developed for the purpose of building exams to

test measures of question difficulty and discrimination; it was built to facilitate improved student performance through a series of question authoring and answering drills. Nonetheless, the extensible structure of the system, including the ease of accessing anonymized versions of the student interactions and the large number of details maintained about the interactions, allows for myriad experiments. Although not designed for exam building, PeerWise is especially well-suited for these experiments and many others in the educational domain.

I describe how I use the PeerWise data for building exams and testing question difficulty and discrimination in Chapter 3. In Chapter 4 I show the results of my experiments using this crowdsourced data from PeerWise and in Chapter 5 I reflect on other applications for this data and for other similar data sets.

2.4 Automatically Measuring Question Difficulty and Discrimination

To measure the usefulness of exam questions, researchers have devised methods for judging both the difficulty of a question and the differentiation power of the answer options [39] [38]. One such approach is Item Analysis Theory [2].

This method for judging question difficulty and item discriminating power relies on models of student performance from the three performance groups previously mentioned. 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 negative number in the "UFN" column in the examples in Appendices G and H.

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, I measure the relationship between the question and the answer option in a way that mirrors student performance as ascertained by Item Analysis.

I use the PeerWise test bank data as example exams to discover what characteristics make a question difficult. This means that the sets of questions that are answered by sets of students are turned from a sparse (and useless) set where there are few shared questions to a very dense exam. There is also additional information in the PeerWise data that I can use for future experiments, such as ratings by the students as they answer questions about how difficult and how good a question was. Future experiments will attempt to link how students performed on a question and how they thought they performed. For example, to compare how discriminating a question was for an instructor to the quality rating given by the students.

Using the PeerWise data, I tried two approaches to create exams that had the most difficult and the most discriminating questions. I viewed this as "exam building" where I want the smallest set of the best questions that told me the most about the students. Both approaches, one based on weighting questions and students by difficulty and performance, and the other based on the most complete set of the same questions answered by the same students can be reviewed in more detail in Appendix H. These papers look at the average difficulty of the questions in the data and test the aforementioned methods to find the most discriminating questions [75] [1].

The field of educational psychology provides methodologies for examiners' test building and taught strategies for students' test-taking. Testing is used in education to gauge whether or not a student has absorbed the information that has been taught. With the successful completion of examinations, educators may move on to new or advanced topics. The unsuccessful results of an exam direct teachers to revisit curricula. In designing exam items, the goal is to create questions "that distinguish between those students that have achieved the learning outcomes being measured and those who have not" [26]. Thus, the ideal test item will be one that enables "the knowledgeable student, and *only* the knowledgeable student, to answer correctly" [26].

In test creation, MCQ exams are called "selection type" because, like true-false and matching question formats, selection type "require[s] the student to choose the answer from among two or more alternatives" [26]. The opposite of "selection type" is "supply type" test items, which include short-answer and essay questions. Of the item types mentioned–MCQ, true-false, matching, short-answer, and essay–all can be classified as objective items, except the essay. Selection and supply type items are also referred to as "recognition" and "recall" items, respectively, in Davis' Educational Measurements and their Interpretation [76].

What is generally identified as a MCQ is called an "item" in educational psychol-

ogy. An item is made up of a "stem" which is in the form of a question or an incomplete statement and several potential answers. The possible answers are called "alternatives" and the incorrect alternatives are called "distractors." The function of distractors "is to distract those students who have not achieved the specific learning outcome being measured by the item" [26]. MCQs are the most generally applicable of the test types. They are used in situations where examiners expect students to "identify" the best answer from a provided set of answer options. According to Gronlund [26], the three primary reasons that MCQs are embraced by teachers are

- "It can be designed to measure a variety of learning outcomes, ranging from simple to complex
- The use of four or five alternates reduces the student's chances of guessing the correct answer
- The use of several plausible incorrect answers for each item provides diagnostic information concerning the most common errors and misunderstandings of low-scoring students."

There are two manners for interpreting the results of a test. The first is called "criterion referenced" and describes directly the student's performance, such as "they completed 80 out of the 100 questions correctly." The second way of interpreting test results is referred to as "norm referenced" and positions the student's results in comparison to those of their peers. An example of norm referenced is "the student performed better than 95 percent of his or her peers." Both types of interpretation are useful in evaluating students, but only norm referenced analysis lets test makers examine the value of *each* of the test distractors because it brings in information on how well the distractors worked across a group of students [2].

"*Item Analysis* means the process of discovering how difficult an item is and how relevant it is to the variables measured in the tests" [76]. Formal item analysis suggests whether each test item does, in fact, help differentiate better students from more average students. The process of formally analyzing items contains a number of steps. Item Analysis consists of measuring both item difficulty, which is shown in step 7, and item discriminating power, which is shown in step 8. Once a cohort (for this example, 100 students) has taken a test containing suitable questions, it is graded. The process for formally analyzing items, adapted from [76] [2] is as follows:

1. Rank the 100 test papers in order from the highest to the lowest score.

2. The set of 100 test papers is split into three groups that represent the top-scoring, middle-scoring, and lowest-scoring students. These three groups are commonly split, lower 27%-middle 46%-upper 27%.

3. The middle set of (46) exams is excluded. These exams are excluded from further analysis because they contain no discriminating information.

4. 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. Figure 2.6 illustrates a sample filled-in template, including all omissions.

5. For each question, the number of responses, including the omissions, should equal the number of students who took the exam, in this case 100. This is shown in Column 6.

6. The number of students in the bottom row, "total," should also equal the number of students who took the exam.

7. In the example, Column 6 shows the cumulative total of responses for each answer option. Item Difficulty can be measured by the number of students who correctly answer the question (in this case, answer option B) divided by the number of students who tried. This result is then multiplied by 100. In the instance shown, the item difficulty is 35%.

8. Item discriminating power represents the positive or negative value of the difference between how many high-scoring students answered a question correctly versus how many low-scoring students did. In Figure 2.6 this would be the positive or negative value of the result of Column 3 subtracted from Column 5. Mitkov et al.'s describes the distractor classes [65] in terms of their usefulness or "UFN."

After measuring item discriminating power, if the resulting number is close to zero, that means that the alternate had low or no discriminating power. If the answer is positive, it means that it was a good distractor because it distracted good students from the correct answer choice [76] [2]. Item Analysis is further discussed in Section 3.3. The need for complete exams to perform Item Analysis motivates the adjacency matrix approach for building exams.

Voorhees discusses what factors play a role in making some questions more difficult than others in the TREC commentary papers [77] [78] [79]. TREC commentary

Column 1 Item Number	Column 2 Choice Letter	Column 3 # in High-Scoring 27%	Column 4 # in Middle-Scoring 46%	Column 5 # in Low-Scoring 27%	Column 6 Total Number
1	A	4	14	5	23
	B	11	18	6	**35**
	C	3	3	2	8
	D	3	5	3	11
	E	5	2	9	16
	OMIT	1	4	2	7
	TOTAL	27	46	27	**100**

Figure 2.6: An example of how to perform Item Analysis including Item Difficulty and Item Discriminating Power. Item Analysis examines the answer and distractor choices that groups of students made for a single question. This figure is adapted from [2].

and perspectives are relevant to this work because I attempt to answer the test set MCQs during the process of using QA approaches for judging the difficulty of a question. I tried two approaches based on previously successful QA techniques. These two approaches are discussed in greater detail in Chapter 3.

I use the information in this background chapter to focus my research for the remainder of this project. In Chapter 3 the methodology for automated MCQ answering using web queries is presented followed by new metrics for analysis of question difficulty and distractor quality. Then I discuss the empirical validity of using student-authored questions, introducing PeerWise and the many potential experiments it supports. Identifying the complexity issues inherent to building new exams out of test banks supports using approximation algorithms. Using approximation algorithms instead of a brute force approach to build exams allows the processing of less constrained types of test banks with this more general approach. PeerWise and exam building will be discussed more in Section 3.4 with links to other background information.

Chapter 3

Research and Analysis

The research conducted and reported in this chapter involved building a system that would not only solve a significant number of MCQs but more importantly could determine question difficulty, suggest discriminating power, and comment on the usefulness of distractors. This chapter begins by discussing my early exploration of this research space and includes further details on the automated system. The components described are involved in the automated system built on a series of steps that take a MCQ in text, translate it into XML, send the answer options to the Web for their definitions, and then compare the returned definitions to the original text of the question. In the following sections I will be presenting the important components of this automated system. The full details may be found in Appendix F.

In the initial system, the comparison method for finding the correct answer was based on a bag-of-words approach and was implemented with ROUGE. The test set allowed me to run experiments building a simple question answering system, but without human performance data, I would not be able to model what makes a question difficult or discriminating. Many researchers run experiments online or in their departments to collect human judgments. To gather human performance data I designed an online experiment that if answered by a sufficient number of students would provide me with enough gold standard data. I designed the experiment interface and planned on using Amazon's Mechanical Turk to manage the experiment. A similar design and the use of Mechanical Turk was also suggested by Mitkov et al. [65]. Recent use of Mechanical Turk to host crowdsourced microtasks has yielded valuable research data [72].

I worked with two psycholinguists with extensive experience in experimental design to build a simple question answering interface. The resulting web pages collected what the test taker thought was the correct answer, but augmented that information

with confidence measures about how difficult they thought other people would find the question. In addition, the psycholinguists suggested using two models, one with a set of answer options and another that showed just one answer option to choose from. The aim of the format was to transition from simply gathering correctness judgments to discovering how difficult the question was. I could use this human judgment data to rate the questions for difficulty and if I used the methods presented in Section 3.3 to create performance cohorts and measure question discrimination power.

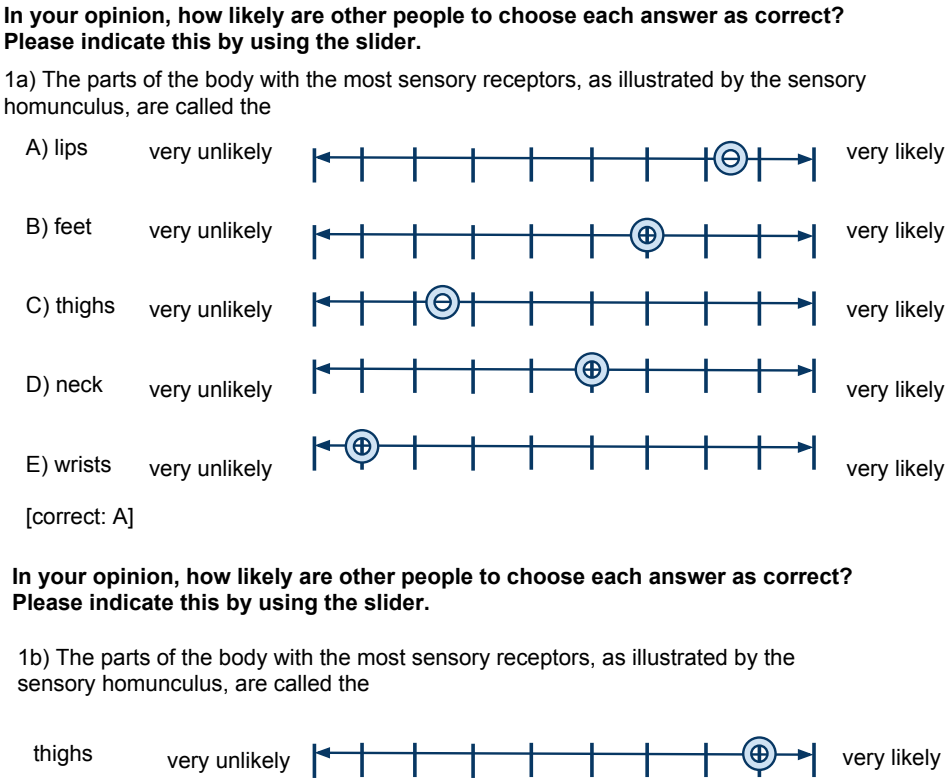

Figure 3.1: This is an example of two questions devised to let humans express their opinions on answer option correctness.

Figure 3.1 shows the proposed human experiment interface. This experiment plays on the intellectual vanity of the test taker. The sliders are set at zero at the start of the experiment, to remove the possibility of "leading" the test taker. The sliders incorporate magnitude estimation, which allows the test taker to individually take his or her confidence regarding the correctness of each question. I had also hoped that the sliders would allow the test takers to provide judgments that were more gradiated that the conventional absolute answer option model. The group of sliders when compared

to one another shows a ranking of the subject's confidence in the most correct to least correct answer option. While this is a valid human experiment design, the PeerWise data sets, (discovered right before these experiments were set to run), provide question difficulty and discriminating power in a more empirical manner.

Now looking at the answer selection aspect of the research, although the ROUGE approach found the correct answer more than two times the success rate as random choice, I decided to improve the hueristic with an LSA-based indexing comparison that was implemented with WordNet and Lucene. I aimed to provide automated question difficulty measures in a manner that coincides with the way that humans judge questions difficult. Section 3.3 goes into detail about the incorporation of Item Analysis into the experiments as a way to measure human performance results by gathering student behavior data in answering questions.

Section 3.1 details early data and experiments, describing my initial attempts on how best to automatically return answers to IDQs and compare these returned results. Section 3.2 built upon those early experiments by creating an automated question answering and difficulty measuring system, which initally used the bag-of-words method to discover the best answers to questions and then moved to a Lucene-based indexing and retrieval method. Section 3.3 describes how I incorporated human results with the automated system to use as a gold standard for measuring the difficulty and discriminating power of the questions. Section 3.4 describes the two approaches used for building exams, matrix-based and weighting-based. These are examined so that the exam data can be best prepared for Item Analysis and resulting question measurements.

3.1 The Query Type Experiments

Collecting MCQ data and building a system that could answer the questions better than if I had answered them at random was a step towards understanding the features of the questions that made them hard. The next step was to look at the results from this first work, using 1000 college-level introduction to Biology and Psychology questions to build a more robust question difficulty measurement system.

There are two research threads in these early experiments. One is focused on collecting the data, called the *test set*, and the second is on the method used for retrieving questions. The questions that I collected were all Inverse Definition Questions. They presented the definition of a biological or psychological term and sought the best match

among a set of options.

The test set data included the correct answer to the questions. Thus, when experiments on the data were run, it could easily be determined how successful the system was overall. However, whether or not the easiest questions or the more difficult ones were being answered correctly was not determined. The question answering system could only answer questions for which the answer was already known. What was the value of this? I used this insight to find a better data set with included human responses for the later questions.

Another aspect of the research project in its initial stages involves discovering the best retrieval method for this problem. Following the lead of current research projects in QA I used the Web to answer MCQs [80]. There are several ways to discover relevant information on the Web so I attempted three query types and recorded how well they correctly answered questions in my test set. I used the best performing method, called "Define: X" in my automated question difficulty system (see more on the Define: X method in Section 3.1.2).

After collecting data and retrieving results, I analyzed those results using a method that has remained the same from the early experiments to those that use PeerWise data. Section 3.2.2 describes the refined comparison method in more detail. Initially, the retrieved question information was compared utilizing a bag-of-words approach and later with LSA-based systems that incorporated WordNet weights into the comparisons.

The test data used and experiments performed in the early experiments include the following:

- DATA: A Test Set of 1000 Biology and Psychology IDMCQs
- RETRIEVAL EXPERIMENTS: Three web queries based on information in the questions
- COMPARISON APPROACH: BOW, not normalized for string length
- GOLD STANDARD: Correct answer known, no human performance data

In the following subsections, test data is covered in more detail and three preliminary web query type experiments are discussed. Note that the weaknesses in these early experiments directed solutions that are present in the automated system are discussed in Section 4.2.2.

3.1.1 The Test Set Data

This section describes the test set used in the early experiments, examines the type of definitional questions used, and describes the limitations of using test sets that do not have human performance data. Further, the definitional groupings that these questions use are reviewed.

A total of 1000 new Inverse Definition, "known as" and "is called" questions and their close variants were found in AP, Regents, CLEF, and SAT practice exam books: [81] [82] [83] [84] [85] [86]. The secondary school exam questions are from two academic subject areas: psychology and biology. The questions were gathered in plain text and there were 440 covering the academic curriculum on psychology and 560 on biology. The questions were located in study support guides either online or in test preparation books, and were used with the consent of the authors. Associated with each question were multiple choice answer options, the correct answer, and often an explanation of why a specific answer was correct. The questions were geared toward secondary school students examined on a government-approved regional or national curriculum.

The patterns used to identify phrases that contain definitions in text were used to filter the question stems added to the test set and later, the PeerWise data. The general form of the questions contained phrases such as "known as," "is called," or varients of these terms deemed semantically similar by a linguist. In these MCQs the stem defines a biological concept, process, or relationship that it seeks the name of. The answer options are noun phrases that could be that term. A more detailed discussion of definitional patterns and motivation on using IDMCQs is included in Appendix C.

3.1.2 The Experimental Design

I explored three types of web queries to answer IDMCQs. All three experiments were conceived as preliminary ways to finding the answer to a definition-oriented MCQ in the science domain. The test set contains MCQs in a form that described a concept, process, or relationship in the answer and listed noun phrases that could be the possible answer. A comparison of what the three experiments query on and the metric for the match is shown in Figure 3.2.

In all of the cases the method for chosing the correct answer is the same: the top bag-of-words overlap measure is used. The three query types (as shown at the top of Figure 3.2) were *A* which sent "Define: X" plus each answer option to the search

engine; *B* which used the question stem as the query text; and *C* which set the text of the question stem and each answer option to be searched upon. Approaches *A* and *B* returned website titles and snippets to compare to the answer options. *C* sought the query with the most web hits, a method successfully employed by Keller et al. [80].

Below the query types in Figure 3.2 is the match metric which details what information was used to decide upon the correct answer. All of the comparisons in the match metric are based on non-stop words. To chose the top answer in *A*, I compared the results of the web query with the text in the stem of the question. The answer option whose returned titles and snippets (often a definition) had the most overlap with the text of the question stem was the best match. In *B* I counted the instances of the terms that made up the answer options in the returned titles and snippets. The answer option that was mentioned most frequently was the best match. In *C* the match was based on the number of web hits associated with the question stem and each answer option. The query with the most web hits was selected as the correct answer.

Query type *B* could be described as "Googling the question" and deciding what the correct answer was based on which of the answer options occurred most frequently in the returned information depending on the query. On the other hand, query type *C* sent each answer option *and* the original question stem to a search engine, iterating for each answer option. The best answer is chosen based on the highest number of web hits.

Initial experiments supported the use of *A* or the "Define: X" method because it was the only approach of all three methods based solely on direct evidence in the question. In experiments *B* and *C* the match was based on indirect evidence. "Define: X" used a search engine shortcut that queries dictionaries, encyclopedias, and documents containing definitions.

Figure 3.3 shows an overview of how the data flows through the automated system. The questions introduced in XML are shown on the left and the database that stores all of the results is shown on the right. The top row of the figure describes the format of the data being presented below the dashed line. Beneath the dashed line the components of the system are (from left to right), the questions, the query methods, the results data, the types of comparison, and the database. The results in the database are in XML and include all of the information gathered in the query process.

Using an XML database allows the nested data to be used in future analysis. The different experiments conducted are shown in the center of Figure 3.3 and clarify that data was used for comparisons with ROUGE and Lucene (further described in Section 3.2). The experiments are shown in Figure 3.2 in the Query Type column and elab-

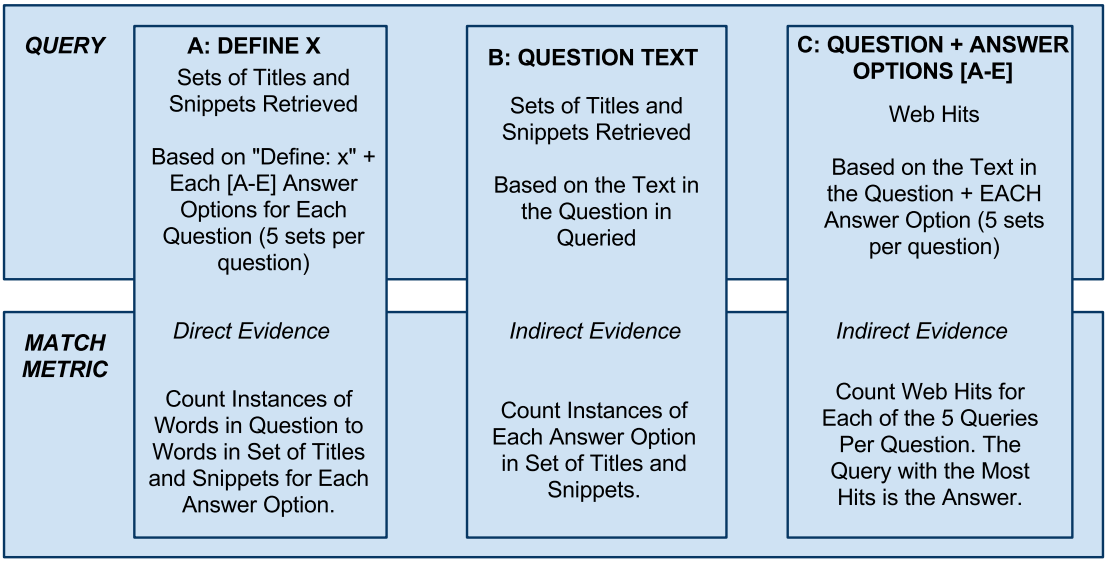

Figure 3.2: The three types of web queries used in the automated experiments.

orated upon in Section 3.1. Greater detail on how this process works is described in Appendix F.

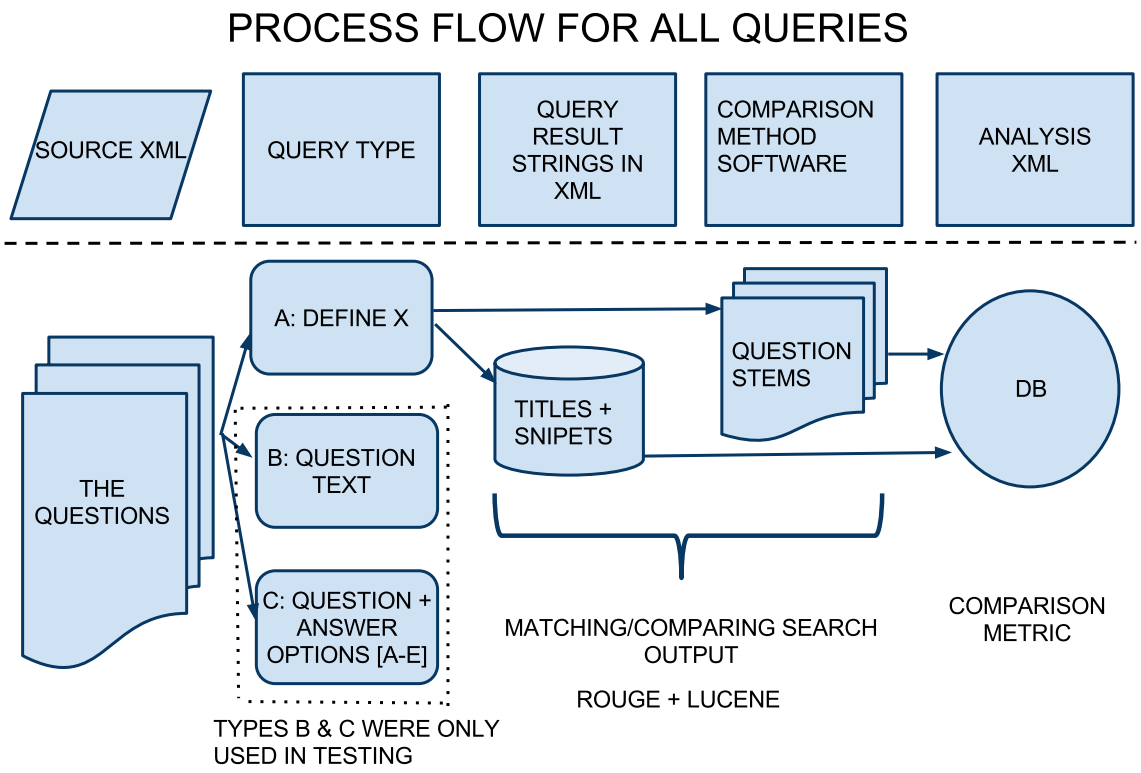

Figure 3.3: This overview shows the question data as it moves from the original XML question on the left through the query and matching steps to the database metric step on the right.

Section 3.1.1 detailed the test set used in the early experiments which were based on using the text in the questions and answer options as search terms. The returned titles and snippets from these searches were then compared to the question stem. The closest match, based on the bag-of-words method, was deemed the correct answer. This is covered further in Section 3.2.2. The early experiments provided several insights that were used in the later experiments. Of the three experiments explored, the "Define: X" format was the most successful at answering these IDMCQs. Limitations with the test set, however, led to looking at another set of questions (from PeerWise) for further research. The need to validate ROUGE's bag-of-words matching algorithm led to my adoption of a Lucene-based approach in the later work.

3.2 The Automated Question Analysis System

This section introduces an overview of the pipeline I built to automatically answer MCQs. The full details of the automated question analysis pipeline may be found in Appendix F. Section 3.2.2 discusses the tools used to compare terms in the returned results to the definitions of the answer options, namely the ROUGE and Lucene software. Then Section 3.3 looks at different metrics for judging human results in MCQ answering.

3.2.1 The Pipeline Overview

There were three main requirements that influenced the development of the fully automated query pipeline. The first requirement was that the pipeline needed to be modularly designed. To move from a manual system to a completely automated one, the new pipeline needed to be designed to enable additional query engines to be incorporated without too much query-engine specific programming. This task was made easier by the Representational State Transfer (REST) [87] architecture, which has been incorporated into the APIs of many popular consumer services on the Web. As REST gained momentum as a straightforward and smart way to interact with the distributed hypermedia of ebsites, it also allowed users to send and receive information from these sites. As a result, the code needed to contact different sites is similar and has aided in design modularity.

The second requirement was to incorporate a flexible comparison system that would allow for building comparison methods incrementally. The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [5] system is an important tool that has been used primarily to compare the quality of the output of summarization systems to a gold standard summary. The gold standard is often human-authored text. Summarization task evaluation includes measuring how close gold standard text strings are to new generated strings. Comparing text strings in the original question to the returned definitions of the answer options is similar to evaluating summarization tasks. ROUGE was a natural choice for this task and will be further described in Section 3.2.2.

The third requirement was to develop a new methodology for handling data as it was retrieved and effectively processing and storing it as it moved through the pipeline. Since the data must be moved through the pipeline in exactly the same manner in each query run, moving the 1000 test questions, the associated query results, and the results of the ROUGE system to an XML database (Marklogic [88]) allowed for a clean and

stable environment for additional system processes.

After the question sets are broken up into uniform groups, the groups are automatically marked up in XML with a Python script. This script also removed any italics or unusual characters that may have caused problems later in the pipeline.

The entire system implementation is shown in Figure 3.4. Each major section of the data flow pipeline is shown left to right starting with the input question documents. The 50 question documents contain 20 questions each and are marked up in XML. These question documents are used as input to the XProc-based query system, which identifies the IDQ and the related answer options, sends the answer options to a search engine with the prefix "Define:," and post-processes the stored results into a version that is input to the ROUGE comparison system [5].

I used Calabash [89], an open source implementation of XProc written in Java. Building an XML pipeline supports data traceability and in the entire pipeline the original retrieved information from the queries was retained. The specific format used by ROUGE was 251 individual documents per question, one for the initial question and 250 for the retrieved result titles and snippets. The output of the ROUGE and Lucene systems were results documents that were then passed to the post-processing component of the pipeline. In this component, the results were merged, filtered, and compared to the actual correct answer. Then, the original XML question document was augmented with this information. Finally, more extensive analysis was provided by using the Marklogic XML database [88].

Once the original questions had been turned into XML documents, those documents became the input to the XProc [90] pipeline which in turn carried out a series of queries and transformations based on the XML tags associated with the data structure of those documents. The input of this system is shown in Figure F.2 in Appendix F. Then, a second XProc pipeline processed the output of these queries into formats accepted by ROUGE (described in Section 3.2.2) and Lucene. These two comparison methods were used to measure the similarity of the question stem to the query text based on the answer options. More detail regarding how these two systems work may be found in the following section.

3.2.2 ROUGE and Lucene

Two systems that compare collections of words from documents to measure the similarity between these documents were used to measure the similarity between questions

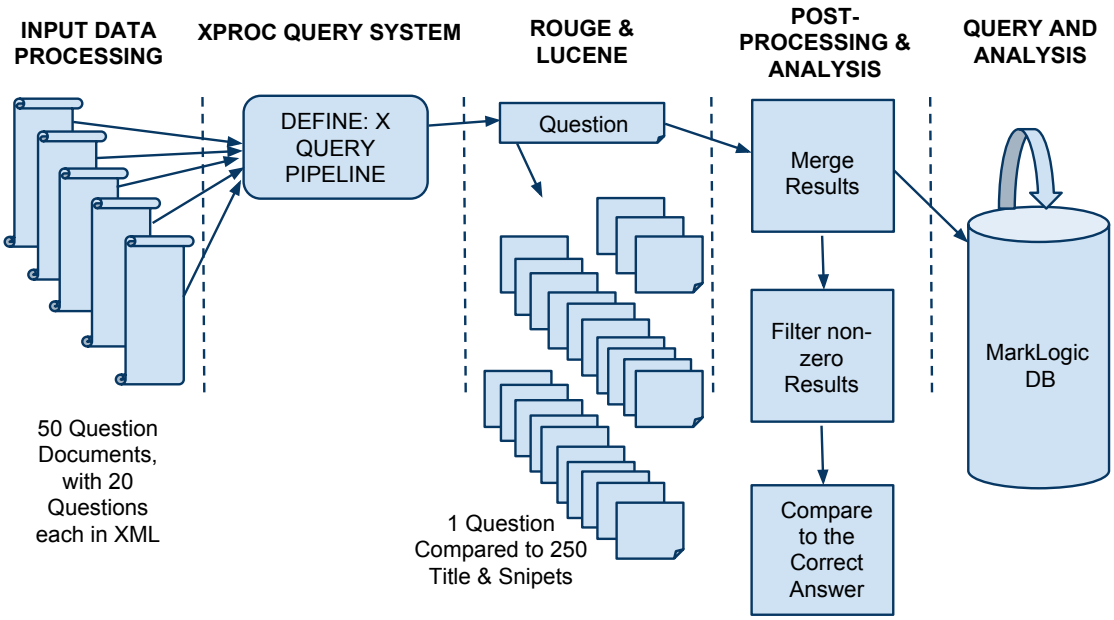

Figure 3.4: An overview of the automated query data flow with the XProc pipeline. Examples of the "input data" to XProc are described in Section 3.2.1. The query types are explained in Section 3.1.2. ROUGE and Lucene are discussed in Section 3.2.2. The results of the system shown on the right hand side of the figure can be found in 4.2.2.

and possible answers. In the arena of MCQs, the answer that has the most similarity to the question stem is selected as the "best" answer. To gather more details on each answer option, they were run though a pipeline (discussed in Section 3.2.1) that returned definitional information from websites as collected in strings of the site's titles and page snippets. This information related to the answer option was compared to the question stem. The answer option that had the most similar text to the question was deemed the correct choice. The two systems used in the comparisons were ROUGE and Lucene. Both are bag-of-words-based comparison methods and to validate my ROUGE results I ran similar comparisons in Lucene. To further validate the simple bag-of-words analysis, I implemented bigrams and a link with WordNet to determine whether these additions could provide better performance for my automated system.

ROUGE (Perl-based) and Lucene (Java-based) are text comparison systems that examine the resemblances between one text string and a set of other possibly matching text strings. Lucene has been ported to several languages and it was implemented in Python for this research. ROUGE and Lucene both judge the similarity of these strings on their statistical word overlap. As a consequence they are appropriate tools for measuring the similarity of inverse definitions to definitions retrieved from the Web. The development of ROUGE was initially motivated by the very expensive cost of human judgments that distinguish good summaries from less viable ones. Human evaluators are used in many tasks in natural language processing and as a result ROUGE, as a replacement tool, has found utility beyond summarization.

ROUGE has also been used in comparing the output of machine translation programs and the task-based evaluation showed results similar to those of humans [91]. In this case, ROUGE aided measuring the similarity between a "candidate translation and a set of reference translations" with matching the Longest Common Subsequence (LCS) and utilizing skip-bigrams [91]. LCS identifies the "longest co-occurring in-sequence n-grams"; skip-bigrams are "any pair of words in their sentence order" [91]. Using ROUGE and Lucene for comparing the output of queries from the Web also benefits from automatically identifying "sentence-level structural similarity" as the text being compared to the original question is a concatenation of the title of the web page and the snippet (or summary) of the content on that page that contains the most relevant information to the initial query [91].

Once the "Define: X" definition queries had been run, the returned titles and snippets may have included the answer option's definition because of the "define" shortcut. This shortcut used the search engine's own metrics to associate the queries specifically

to strings of text that make up definitions. Thus, the retrieved title and snippet were used to assess whether the retrieved definitions were approximations of the inverse definition.

ROUGE and Lucene are flexible tools that allow different metrics for comparing strings. The baseline ROUGE implementation used in this research was ROUGE 1, which sought the unigram match of the literal word overlap between the original IDQ text and the text of the retrieved title and snippet pairs. The ROUGE 1 implementation also omitted stop words and stemmed all of the remaining content words using the Porter stemmer. Many of the IDQ contain either the word "known" ("is known as") or "called" ("is called"). "Known" is on the stop word list, but "called" is not.

Example 6

The process that releases energy for use by the cell is known as

A. Photosynthesis
B. Aerobic metabolism
C. Anaerobic metabolism
D. **Cellular respiration (correct answer)**
E. Anabolism

Figure 3.5 shows the ROUGE scores for the text strings from Example 6 that are presented in Figure F.8 in Appendix F, except for the numbers that had zero scores. The first number in the line, in this case "151," corresponds to the number of the retrieved result being compared to the model_path, or original question. As detailed in Figure F.7 in Appendix F, this means that numbers 151, 152, and 153 are the first three results for answer option "cellular respiration" and number 162 corresponds to the twelfth retrieved result. The version of ROUGE used in the comparison is listed after the retrieved result number, followed by the average recall ("Average_R"), average precision ("Average_P"), or average F-score ("Average_F") and the numerical score with a confidence interval ("conf. int."). The lines in between the dotted line and the dashed line summarize all of the per question results in one line.

For number 151, the bigram scores were 0, revealing that while there was some individual word overlap, there were no overlapping two-word phrases. While the answer option "cellular respiration" was the correct answer to the question, retrieved results number 154, 155, and 156 had no overlap with the text of the original question and as a result, are not shown in Figure 3.5. Also, result numbers 152 and 153 have no overlap when using the bigram comparison.

```
151 ROUGE-1 Average_R: 0.25000 (95\%-conf.int. 0.25000 - 0.25000)
151 ROUGE-1 Average_P: 0.05263 (95\%-conf.int. 0.05263 - 0.05263)
151 ROUGE-1 Average_F: 0.08695 (95\%-conf.int. 0.08695 - 0.08695)
.........................................
151 ROUGE-1 Eval 1.151 R:0.25000 P:0.05263 F:0.0869
-----------------------------------------
152 ROUGE-1 Average_R: 0.25000 (95\%-conf.int. 0.25000 - 0.25000)
152 ROUGE-1 Average_P: 0.06667 (95\%-conf.int. 0.06667 - 0.06667)
152 ROUGE-1 Average_F: 0.10527 (95\%-conf.int. 0.10527 - 0.10527)
.........................................
152 ROUGE-1 Eval 1.152 R:0.25000 P:0.06667 F:0.10527
-----------------------------------------
153 ROUGE-1 Average_R: 0.75000 (95\%-conf.int. 0.75000 - 0.75000)
153 ROUGE-1 Average_P: 0.20000 (95\%-conf.int. 0.20000 - 0.20000)
153 ROUGE-1 Average_F: 0.31579 (95\%-conf.int. 0.31579 - 0.31579)
.........................................
153 ROUGE-1 Eval 1.153 R:0.75000 P:0.20000 F:0.31579
-----------------------------------------
162 ROUGE-1 Average_R: 1.00000 (95\%-conf.int. 1.00000 - 1.00000)
162 ROUGE-1 Average_P: 0.23529 (95\%-conf.int. 0.23529 - 0.23529)
162 ROUGE-1 Average_F: 0.38095 (95\%-conf.int. 0.38095 - 0.38095)
.........................................
162 ROUGE-1 Eval 1.162 R:1.00000 P:0.23529 F:0.38095
-----------------------------------------
```

Figure 3.5: ROUGE 1 output for the first three and twelfth retrieved results for "Define: Cellular Respiration."

The precision score in Figure 3.5 corresponds to the original question containing four words that overlap with the 17 content words in result D_12. The recall is 1.0 because all of the words in the question text occur in the retrieved title and snippet string. The F-score, calculated below, is also known as accuracy and is the metric used to judge the top results of the ROUGE module:

$$F = 2 \times \frac{precision \times recall}{precision + recall}.$$

Question 162 has the highest score due to the overlap of the terms "process," "releases," "energy," and "cell." Figure 3.6 presents the inverse definition text and text that ROUGE considers for comparison in three steps. Step one reiterates what is being compared: the text in the question stem to the retrieved results from "Define: X" and each of the answer options. Step two shows what the returned text is for answer option D, "Cellular Respiration." Step three shows the actual text that is compared, the content words. In Figure 3.6, stop words are shown crossed out, and the remaining words are in plain text.

While the ROUGE comparisons were used as a baseline method, the comparisons created using Lucene were implemented to validate the bag-of-words approach and extend it with more types of matches. Using Lucene, the matches increased to include bigrams and WordNet weights. Bigrams extend bag-of-words methods by seeking word pairs that co-occur in both documents being compared. Introducing WordNet allowed more variants of the terms in the comparison to be included.

The experiments performed using Lucene were organized with some additional parameters in an effort to pursue improvements in the matching algorithm. Two methods were used for building the search index. They were either *individual* or *grouped*. Individual meant that only the title and snippet for each of the 50 queries per answer option were used as a comparison document to the orginal answer stem. Grouped meant that all 50 titles and snippets were used as one document for comparison.

The answer match metric was also expanded from matching the word overlap as used in ROUGE, to include two other measures for choosing the best answer. The ROUGE system found its answer choice by averaging the top 10 results from each set of web queries and choosing the answer option with the highest score. In Lucene, that approach was also used, and called the "aggregate" method, but I also incorporated two other choosing metrics: top highest score and top number of hits.

The top score is rather straightforward: choose the answer option that has, in its set of returned strings, the single highest similarity score with the question stem. In

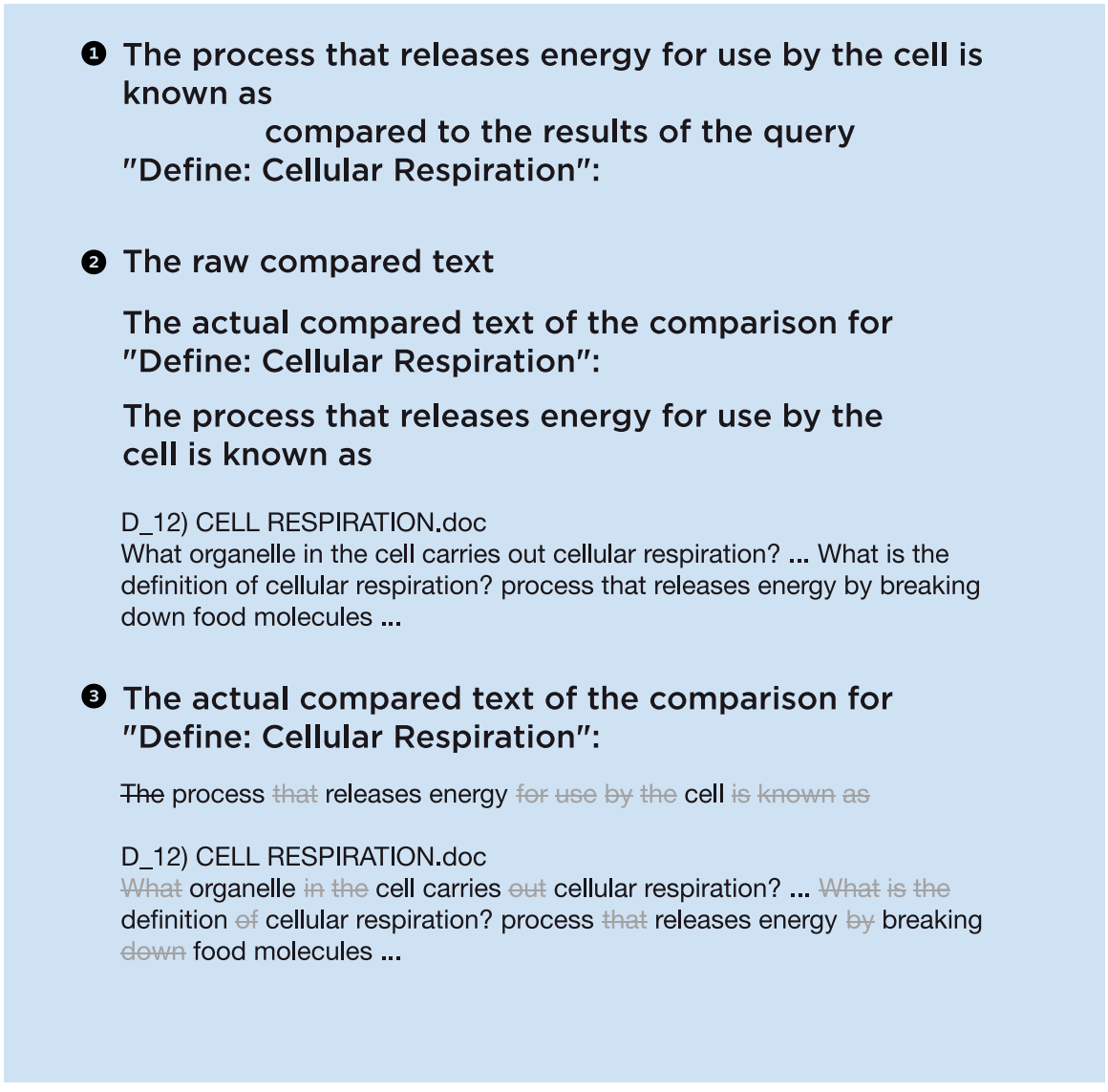

Figure 3.6: The Top ROUGE Score Comparison for "Define: Cellular Respiration."

the top number of hits method, the best answer option was chosen based on the most occurrences of the words in the question stem in the answer set(s). The two indexing methods and three matching metrics provided results that could be used as both validation of the ROUGE system and iterative improvements on it, as well. Similarly, introducing WordNet word weights allowed the matching to occur on terms that did not occur in the original terms being compared, but expanded the possible matching set. In these experiments, WordNet was used to expand the question stem. Adding bigrams and WordNet information increased the correct matching of the MCQ answer options significantly. The results of these experiments may be found in Section 4.2.1 and 4.2.2.

3.3 The Analysis of Human Test Data

This section reviews how human performance data is used to model the difficulty and discrimination of a MCQ. Following a discussion of what factors make a question difficult there is a review of what makes a question discriminating (Section 3.3.1). Then, in Section 3.3.2 PeerWise data is recognized as a resource that contributes the detailed human results permitting Item Analysis and performance modeling. Next, in Section 3.4 two methods for building exams are introduced. These exams are then evaluated with Item Analysis to measure their question quality.

3.3.1 Analysis of MCQ Difficulty and Discrimination

Based on the research of Mikov et al. [65] and Gronlund [36], my research adopts terminology and methods for analyzing test item usefulness, discrimination, and difficulty. The analyzsis of questions based on how students perform in a group with the goal of improving exam-based MCQs is broadly called Test Item Difficulty. Within the study of Test Item Difficulty are several measures including Item Analysis, which evaluate how each student performed based on which answer choice he or she made. This section explains how Item Analysis and its components are used to analyze the PeerWise data. The results and discussion of this analysis may be found in Section 4.3.

A question may be difficult in many ways. The stem may be confusing or poorly scoped. The topic of the stem may be from an obscure corner of a discipline or use ambiguous terminology. Further, when a question has multiple answer options, high-quality incorrect options and discriminating distractors can make a question difficult.

One goal of this work is to discover characteristics of difficult and discriminating questions and use these traits to support an automated approach to assessing which questions are most valuable in determining the comprehension levels in a group of students. Since MCQs are used to measure the comprehension level of students, these questions cannot be so difficult as to be impossible for the best students to answer. MCQs aim to differentiate the low, middle, and high-performing cohorts of students by their performance.

To measure question difficulty, researchers have devised several methods including Item Analysis for judging both the difficulty of the question and the differentiation power of the answer options [2]. Item Analysis was introduced in Section 2.4 and the process is briefly reviewed here. Section 4.3.2, shows the results of running questions through Item Analysis. Appendices G and H present more extensive Item Analysis results on a the PeerWise data from Courses 1 and 2.

Here is an example of my approach. Assume a class of 100 normally distributed students takes an exam. It is graded and the exams are ranked from the highest score to the lowest. The 27 exams with the highest score and the 27 exams with the lowest score are selected. These exams represent the best-performing and worst-performing sets of students on the test. The 46 exams in the middle are the average-performing students, and are discarded. To summarize my approach:

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. Figure 2.6 illustrates a sample filled-in template, including all omissions.

2. *Item Difficulty* is measured by the percentage of students who answered a question correctly. The lower the percentage, the more difficult the question is. In Figure 2.6, the correct answer is B and the question has an item difficulty of 35%, as shown in column 6. In general, if more than half of a class got a question incorrect it is considered difficult, but difficulty for grading purposes is often interpreted on a curve.

3. *Item Discriminating Power* is the difference between the number of high-scoring students and the number of low-scoring students who chose the same correct answer option divided by half of the number of students who were included in Item Analysis. This means that this measure is normalized for all the students in the study, not just the ones answering this particular question. "An item with no

discriminating power would be one where an equal number of pupils in both the upper and lower groups got the item right" [2].

4. *Distractor Usefulness* is the term coined by Mitkov et al. [65] and it is based on "item effectiveness" which was a method used by Gronlund [36]. Usefulness is measured by comparing the number of students in the top-performing group who selected an incorrect answer option to those students in the lower-performing group choosing that same incorrect option. This is recorded for each distractor and is not normalized. There are situations where almost equal numbers of high and low-performing students chose the same wrong answer, and that distractor is considered a *poor distractor*. A distractor chosen by no student is called a *not useful* distractor.

Gronlund introduced one method for measuring item discriminating power, where D = the Index of Discriminating Power; Ru is the number of students in the high-performing cohort who answered the question correctly; Rl is the number of students in the lower group who answered the question correctly; and 1/2 T is one half of the total number of students included in the Item Analysis. This total number of students includes students who omitted questions. Gronlund's approach gives the Index of Discriminating Power, D, for the correct answer, but does not address the possible relationship of the other answer options to the discriminating power of a question [36]:

$$D = \frac{Ru - Rl}{T/2}$$

For example, in Figure 2.6 the correct answer is B. B was the favorite answer choice for the high-scoring student cohort with 11 students choosing it. In contrast, B was also the favorite answer choice of the low-scoring students and was chosen 6 times. In this case, the Item Discriminating Power would be (11-6)=5/50=.1.

Another feature of Item Analysis is how distractors are characterized in terms of their contribution to the utility of the question. The goal of distractors is to trick the lower-performing students into chosing them instead of the correct answer. Thus, "usefulness" measures group these distractors into three classes: poor, not useful, and useful.

A "poor distractor" is a positive number and indicates that more high-scoring students found this incorrect answer attractive than did lower-scoring students. Multiple poor distractors in a question correspond to more than one incorrect answer option being attractive to higher-performing students. One hypothesis for what makes a question difficult is that the question has multiple poor distractors. This means that good

students split their choices between several alternates and that these choices are not considered good question options by lower-performing students.

A "useful distractor," a negative number, catches more low-scoring students than higher performing ones, and is the aim of question authors. When the number of higher- and lower-performing students who chose the same answer option is equal, that is a "not useful" distractor. When few or no students chose an incorrect answer option, it is "not useful."

A high-scoring student is one who answers most questions correctly, but when his or her answers are incorrect, chooses the best distractors. The best distractors are the choices most likely to be good alternates to the correct answer option. 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. That cohort will behave like low-scoring students, with a near equal spread of multiple distractors being chosen.

3.3.2 Determining Difficult and Discriminating Questions

Many NLP-based exam generation systems rely heavily on previously produced (authored by human experts) real exam data to refine how similar the distractors need to be to the correct answer to be classified as "good" [64]. In the case of standardized comprehension or aptitude exams, this means having access to sets of exam data, which include the questions and detailed, question-by-question results from thousands of students. Unfortunately, such ideal data is very difficult to obtain.

I procured data for two sets of MCQs from university-level introductory biology classes using the PeerWise question creation system [10]. Introduced in Section 2.3 PeerWise consists of questions that are created by students and answered by their classmates. Instructors can review the questions or use some of the better questions for future exams. Since answering these questions may not be compulsory, the resulting data is a set of questions that have been answered by students but not all of the questions have been answered by the same students.

The process of choosing questions for the experiments consisted of automatically collecting the subset of questions that used inverse definition constructions such as "is called," "known as," and "is named." Inverse definition questions describe a term or process by providing a definition and seek the name of the process. This question format is frequently used in the sciences where mastering domain-specific concepts

are a key measure of comprehension.

Further filtering of the questions removed any queries that contained or were structured with images, symbols, true-false, analogies, or negation. Questions with three, four or five distractors were sought, but four and five distractors were the majority as only 4 three-distractor questions were included. This preprocessing reduced the data Set1 of 752 initial biology questions down to 148. (Future research could analyze these types of questions but they are outside the scope of this effort.) To reiterate, the questions contained:

- no images
- no true/false or yes/no
- no mathematics needed to solve the question, no numbers, and no symbols
- no negation, exceptions, comparisons, or superlatives
- no why, how, or explanation-oriented questions
- no answers that consisted of multiple terms or groups
- no "all of the above" answers
- no analogies

Next, the question sets were manually reviewed to check for obvious spelling and grammatical errors. There were fewer than a dozen spelling and grammatical errors. Considering the thousands of questions initially reviewed in building these data sets and the hundreds that were further reviewed once they passed through the automated filters that specifically sought inverse definition questions, this result suggests that the interface fosters effective self-policing of the question content. The interface for rating and commenting on the question is shown in Figure 2.5 in Chapter 2.

Then all of the additional information associated with the questions, such as the related question materials, was collected for preprocessing. These materials consisted of the unique question id, the timestamp of when the question was taken, the unique student id, the average rating (0 to 5), the average difficulty (1 to 3), the total number of responses, the total number of ratings, the correct answer, the number of answer options, the text of the question, the text of the answer options, and an explanation, if one was present.

A database output example of a PeerWise question can be found in Figure 3.7. In Figure 3.7, the fields are, from left to right: the unique question identifier (31522), the timestamp, the unique and anonymous identifier of the student who authored the question (11713), the average "goodness" rating (2.7742), the average difficulty (1.0000), the number of times the question was answered (244), the number of times the question was rated (62), the correct answer choice (D), and the number of answer options (5). Next is the question stem, the answer options, and an explanation of why the correct answer was the best answer. In this case, the author gave definitions of the terms listed as answer options. Any tags that denote emphasis such as bolding or italics remain from the authoring environment. For example, in Figure 3.7, the "strong" tag is used by the author for emphasis.

The PeerWise data's transition from GUI to plain text to database consisted of three steps. First, I stored all of the additional information on the question, including when it was taken and who authored it, in a MySQL database. The questions were filtered based on the noted constraints and the output per question is human readable, as seen in Figure 3.7, with each information column-delimited.

```
|
| 31522 | 2010-03-22 00:13:23 | 11713 |     2.7742 |        1.0000 |            244 |           62 | D      |
5 | <p>What is the name of the area's between osteon's?</p>
| <p>canaliculi</p>
| <p>lacunae</p>
| <p>lamellae</p>
| <p>interstitial lamellae</p>
| <p>Volkmann's canals</p>
| <p><strong>Canaliculi - </strong>small channels that run through the ECM allowing the flow of fluid and
blood.</p><p><strong>Lacunae - </strong>lake like structure where osteocites live (connected via canaliculi)</
p><p><strong>Lamellae - </strong>concentric, cylinder-shaped layers of calcified matrix. </
p><p><strong>Volkmann's Canals - </strong>oblique channels that connect osteons to each other and the
periosteum</p><p> </p><p><strong>Interstitial lamellae - </strong>islands of calcified matrix between
osteons.</p><p> </p><p>pg 233 (anatomay and physiology 6th edition)</p>
|
```

Figure 3.7: Column-delimited version of the data gathered from the PeerWise GUI output from a MySQL database.

Then, I reprocessed the plain text question materials into an xml structure that the pipeline system discussed in Section 3.2.1 could take as input. Finally, I gathered the data on the students who answered the questions that met the IDQ criteria from Paul Denny of the University of Auckland, creator of PeerWise. Denny provided a list of unique, anonymized student ids representing each student who answered the IDQ questions. These ids were consistent across each question set. Each data set had five associated files of information –ratings, concept or topic tags, comments for all the questions by each student, each question-answer result by student, and all of the

initial question materials– which were combined into one database. An example of the ratings data is shown in Figure 3.8.

id	timestamp	user	question_id	difficulty	score
234151	2010-03-02 17:13:42	10651	25953	1	5
234191	2010-03-02 18:21:21	10629	25953	0	2
234286	2010-03-02 19:18:01	10677	25953	1	4
234291	2010-03-02 19:18:59	10680	25953	1	5
234319	2010-03-02 19:45:18	10690	25953	1	4
234332	2010-03-02 19:48:58	10603	25953	1	4
234552	2010-03-02 20:44:24	10713	25953	0	3
234587	2010-03-02 20:51:21	10718	25953	0	4
234616	2010-03-02 20:57:37	10714	25953	0	3
234890	2010-03-02 22:01:52	10741	25953	1	3
235120	2010-03-02 22:55:38	10751	25953	0	3
235556	2010-03-03 02:33:00	10818	25953	0	3
235779	2010-03-03 09:05:39	10836	25953	0	4
236062	2010-03-03 10:37:11	10646	25953	0	3
236132	2010-03-03 11:02:17	10863	25953	0	3
237176	2010-03-03 15:57:37	10342	25953	1	3
237494	2010-03-03 17:05:13	11004	25953	0	2
238195	2010-03-03 19:47:11	11049	25953	1	2
239509	2010-03-04 09:45:11	11148	25953	0	3
239700	2010-03-04 10:52:10	10610	25953	1	3
242461	2010-03-05 12:15:20	11288	25953	2	4
242772	2010-03-05 14:21:22	11106	25953	1	4

Figure 3.8: Plain text version of ratings data with the column headers from left to right: the unique identifiers for the instance of a question being asked, a timestamp of when the question was answered, the unique user id of who answered the question, the question id, the difficulty rating given by the student, and the "goodness" or quality score given by this student.

One potential research problem is the hypothesis that PeerWise attracts the better performing students to practice and build their expertise in a field. The better students may tend to both author and answer more questions than their lower-performing peers. Thus, the PeerWise system may skew Item Analysis from a more conventional normal distribution across performance cohorts to a tight cluster of top-scoring students versus a long tail of the middle- and lowest-performing students. This hypothesis was tested and revealed that a sufficiently large group of potential students participated in the exam for meaningful Item Analysis [75]. These results are presented in greater detail in Section 4.3.1. Consequently, the bell curve comprises the three performance cohorts.

The lowest-, middle-, and highest-scoring students, did shift to the left, but the three cohorts were distinctly discriminated.

The PeerWise data has an exhaustive amount of detailed information that covers the authoring and taking of course-related questions. Much of this information is outside the scope of analyzing question difficulty and discriminating power. There is sufficient additional question information for myriad other research projects, but there are a few characteristics of the data in general that reflect its value for this and additional research. These characteristics are as follows:

- Data Set1 comprised 1055 students and 148 questions split into two subgroups.
- Data Set2 comprised 887 students and 132 questions split into two subgroups.
- The fewest number of questions answered by any of the students was 1. 101 students in Set1 answered only 1 question; 152 students in Set2 answered only 1 question.
- 112 was the most questions answered by a student in Set1; 11 students answered 112.
- The average number of questions answered by students in Set1 was 26.6 and in Set2 was 35.8.
- None of the students answered any question more than once.
- None of the questions were so easy that all of the students answered them correctly, nor so hard that none of them got them correct.
- Set1 contained 31,019 answers and Set2 had 31,314 answers.
- The most times a single question was answered in Set1 is 439; the least was 89.
- In Set2, 331 was the maximum number of times a question was answered; 132 was the minimum.
- There were 62,333 distinct answers or questions that were answered in total.
- There were 20,532 incorrect answers of the total 62,333 answers, or 32.7%.
- There were 14,094 question ratings, and each of the 280 total questions from both sets were rated at least once.

The MySQL queries used to gather this information about the data are shown in Appendix E. The numbers shown in the list above reveal that indeed, many students were answering the questions and there was definitely a group who was answering nearly all of the questions provided. In the two courses that produced this PeerWise data, student participation was voluntary and did not affect the student's grade. Having a set of talented, involved students does not invalidate or skew this data but rather shows that the participants mirror other classes of students.

3.4 Two Approaches for Building Exams [1]

This section describes the two approaches used to build exams so that Item Analysis may be performed on the resulting exam-like data. Building exams and performing Item Analysis on questions creates human performance data that may be used to model student performance on other questions. First, a graph-based representation is presented for gathering training data from existing web-based resources. Then, a complementary method is presented which is based on weighting questions by difficulty for building an exam. Further, using Item Analysis Theory [2], these newly created exams are analyzed and both the item difficulty and the discriminating power of the questions are measured. These measures suggest characteristics that can be used by an automated question analysis system for rating question difficulty and discrimination.

Then, a method is presented that efficiently builds new exams that consist of only these discriminating questions and demonstrates the effectiveness of this new set of questions by monitoring student performance group movement across exams of different sizes. This supports the determination of an optimal size and question difficulty-level for an exam to achieve maximum subject discrimination. Finally, there is a discussion regarding how this is but one application dealing with human judgment data. Numerous fields of research and applications will benefit from these techniques and ideally this stimulates collaborative ideation in the crowdsourcing community.

3.4.1 Matrix-Based Approach

My approach for representing the students and questions in the test bank is with a graph: An "exam," where every student answers every question, would be a complete bipartite graph (or biclique). I am seeking a good set that is similar to an exam: I am seeking the students who have answered the most questions in common. Turn-

ing a set of questions that have been answered by some students into an exam is an NP-hard problem. This problem is described in theoretical detail by [92] [93] and is linked to the process of exam building presented in Appendix D. There are other equivalent methods for addressing this problem, including Markov Chain Monte Carlo algorithms. After considering several of these roughly equivalent approaches, I chose the adjacency matrix approach and later evaluated that approach against a question and student weighting-based method. More background information may be found in Appendix D (with implementation results in Appendices G and H), but in the following section is a brief description on how to build an "exam," where every student answers every question as represented by a complete bipartitie graph (or biclique) [94].

Given an incidence matrix M of students and questions, where the rows of M correspond to students and the columns correspond to questions, I can generate covariance matrices S and Q, as seen in Figure D.5 in Appendix D. S is defined as $M \times M^T$, which generates a covariance matrix where S_{ij} shows in how many students' questions student i has answered in common with student j. Transposition is the interchange of row i with column i. Q is defined as $M^T \times M$, which generates a covariance matrix where Q_{ij} shows how many students have answered question i as well as question j. This can be seen graphically in Figure D.2 in Appendix D. S and Q were then used heuristically to compute a sufficiently large clique of questions that have all been answered by the same set of students.

The steps for building and sorting the covariance matrices are as follows:

1. Collect the data in triples of student ID, question ID, and answer choice.

2. Order the students by the number of questions they answered.

3. Build the incidence matrix M, with students corresponding to rows and the questions to columns. If a student answered a question, a 1 is placed in the appropriate column, if they did not, a 0 is placed in the space. The incidence matrix in Figure D.6 (1) is the bipartite graph shown in Figure D.2 in Appendix D.

4. Compute $S = M \times M^T$. A heat map of S can be seen in Figure D.3 in Appendix D.

5. Compute $Q = M^T \times M$.

6. Find the most correlated students by computing the vector s by summing over the rows of S, thus, $s = \sum_j S_{ij}$. Sort the rows and columns of S based on the

ordering of s because S is symmetric. This effect can be seen in Figure D.1.

7. As above, find the most correlated students by computing the vector $q = \sum_i Q_{ij}$. Sort the rows and columns of the matrix Q based on the ordering of q.

This sorting process provides an heuristic for selecting highly correlated students and questions and the results are reflected in the heat map in Figure D.3 in Appendix D.

3.4.2 Question Weighting-Based Methodology

In the question weighting-based methodology, I take a different approach than the clique-based method which 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 and they are given a score. 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 eliminate them from the question 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 reveal the variations in the comprehension levels within the larger group of students.

Finding the boundaries between too easy, too hard, and discriminating questions 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 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. This method attempts to correct

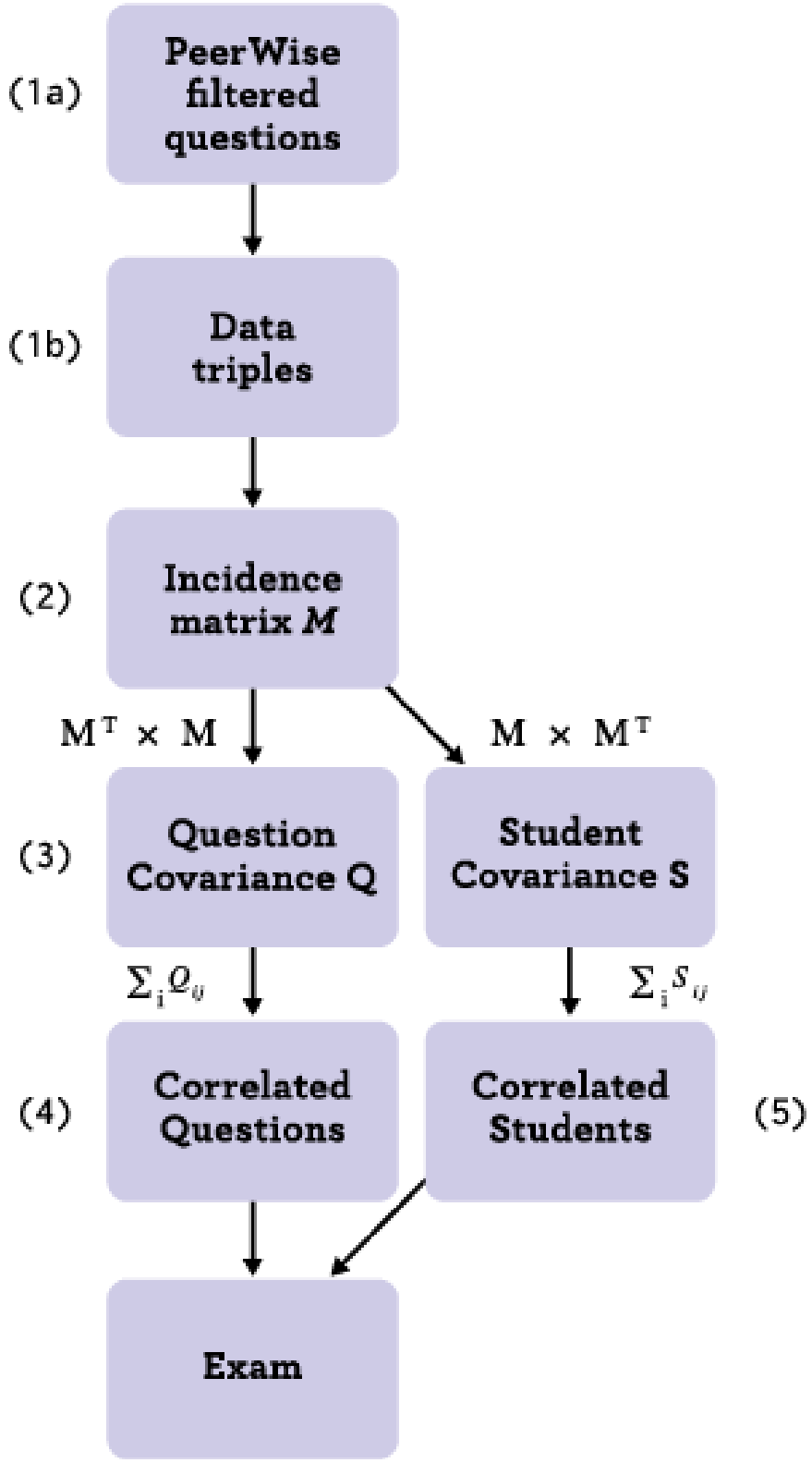

Figure 3.9: In the clique-based approach, an exam is created using the steps for building and sorting covariance matrices.

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 was created who have answered at least three questions. This eliminated data from students who have answered fewer than three questions which is the minimum needed to separate a group of students into performance cohorts (high-, middle-, and low-performing students). A "weight" vector, **w**, was created where each element of the vector represents the weight for a question. The questions were 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 by any student, and a weight of 1 is given to a question that was always answered correctly. Components of the weight vector are calculated using:

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

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. The distribution of question weights can be seen in Figure 3.10. Weights are in the range [0,1] where weights closer to 0 correspond to very difficult questions and weights closer to 1 correspond to very easy questions. The goal is to find the middle band of discriminating questions that are neither too easy, nor too difficult.

This information was then used to rate the students by how they performed on the questions. The weighting-based methodology grouped the types of questions students answered, whether they were easy or hard. These measures were tested by grading every student's answer to every question that had at least 3 students answer it. That would constitute the smallest viable exam that could be split into 3 cohorts, with each student in one cohort.

Figure 3.10 shows the question weights for both Course 1 and Course 2. In general, 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. The questions in Course 2 had a similar difficulty distribution to those in course 1 but did have more questions that were easier than those in course 1. This is shown in Figure 3.10 where the line for Course 2 results

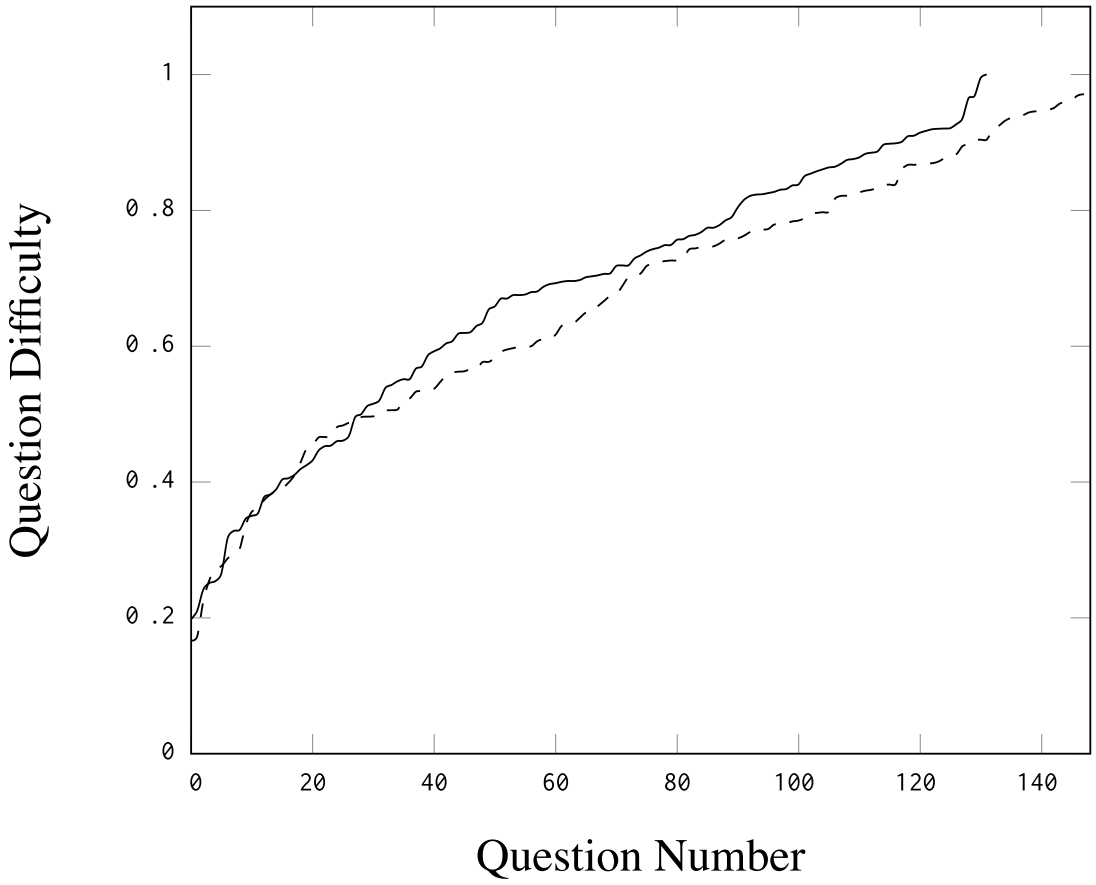

Figure 3.10: Question weights in Course 1 and Course 2 where lower values correspond to questions of higher difficulty. The dashed line indicates weights for Course 1. Course 1 and 2 had 148 and 132 questions, respectively.

deviates from the course 1 results for questions with a difficulty measure of 50% or higher.

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. To score the *i*th student m is the number of questions that the student has attempted to answer and $c(i,x)$ is the correctness of the 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:

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

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 presented in Section 3.4.1.

Next, I 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. This operation was performed by sorting question weights and making 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. This movement can be seen in Figure 4.5. 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.

Chapter 3 presented the methodolgical motivation for the research experiments in this work. Chapter 4 refers to the approaches described here while presenting the results and a discussion about what those results mean. Section 3.1 reviewed the series of early query type experiments and Section 4.1 reports on the results that directed future experimental design. In Sections 3.2 and 3.3 the automated system's comparison metrics and human data processing method were introduced. The corresponding results and discussion is found in Sections 4.2 and 4.3. In Section 3.4 two methods were presented for building exams and Section 4.4 contains the results of this exam building.

Chapter 4

Research Results and Discussion

In this chapter the results of the automated system and human difficulty measuring methods described in Chapter 3 are presented. Section 4.1 presents the results from the query type experiments. Section 4.2 describes the results from the bag-of-words and Lucene matching metrics and explains how the system worked on the more difficult and discriminating questions. Section 4.3 shows the results after processing the questions using Item Analysis and details not only the difficulty and discrimination measures of the questions but also the usefulness of the distractors. Section 4.4 covers the results of building exams from human performance data. Then, the results are reviewed of both the weighting-based and matrix-based methods for building exams.

4.1 The Test Set Results and Analysis

This section reviews the results of the three query type experiments and then presents a discussion of their automated analysis. The data used in these experiments are described in Section 3.1.1 and consist of 1000 MCQs. Concerns that question stem size as well as answer option length might have a confounding effect in the results motivated the analysis in Section 4.1.1. (Basically, it was concluded that a larger bag size for the stem might support more correct answers.) This effect is considered for the full data set in Section 4.2 but was also addressed during the query type experiments presented in this section.

4.1.1 Results and Discussion

The query type experiment used sampling methods, where the sample size was justified according to appropriate statistical criteria [95]. Random sampling selected questions representing the three stem size bags present in the question set. There were questions with 5 and fewer content words, 6 to 9 content words, and 10 or more content words.

Table 4.1: Results controlling for stem length from the query type experiments.

Query Type	*A* "Define: X"	*B* Query stem	*C* Web hits
Total Percent Correct	57%	61%	44%
Correct with stems 10 and greater	100%	66%	33%
Correct with stems 6-9 words	40%	100%	66%
Correct with stems 5 and fewer	33%	17%	33%

There are two questions that arose from the early data analysis: what method for answering inverse definition questions best controlled for stem word bag size and what method best controlled for answer option length? The concern was a larger stem bag size might support more correct answers. Similarly, the way that the answer options were compared to the results sets in experiments *A* and *B* might also have a confounding effect on the results.

As Table 4.1 shows, questions with 6 or more words in the stem were answered correctly more often than those with shorter stems. This suggests that the greater the amount of information provided in the query, the more likely the chance of getting the answer correct in the three types of experiments performed.

Close analysis of the question answer options for concept identification showed three answer types: 1 word, 2 words (usually a colocation), and more than 2 words. I randomly selected questions based on these answer alternatives and tested how the different query types performed on these classes of answers. The results are shown in Table 4.2.

The nature of inverse definition questions is to present a definition and seek the concept being defined. In the biology domain, the answer options are concepts and they are primarily one word terms or two word noun phrases that colocate or are paired to indicate one concept. There are examples of longer answer options, and each of the query types performed differently depending on the comparison format. In two of the three experiments, the web query with the most word overlap with the answer options was used as the "choice" of the system. In the third experiment the most web hits were

used to choose the top answer for each question.

Table 4.2: Results controlling for answer option length from the query type experiments.

Answer Option Length	1 word	2 words	> 2 words
Query Type *A* "Define: X"	83%	50%	100%
Query Type *B* Query stem	83%	50%	none
Query Type *C* Web hits	none	100%	none

Table 4.2 shows that query type *A* performed best with varied answer option length. This may be related to the underlying query methodology that uses the "Define: X" shortcut to search specifically for definitions. Thus, multi-word definitions are maintained as colocated words when they are sent to the Web for results. Concepts such as "active transport" are defined as a unit and not individually with the definition of "active" and "transport" combined. The returned definitions were compared to the question stem and this approach does moderately well when stem length is controlled.

In query type *B*, the method for discovering the best answer was to count the word overlap between the web search results of the question stem, and each of the answer options. The terms in the answer options were compared to the web data individually and each instance of one of the words (if more than 1) was counted as a proportional hit, in cases where there was a tie between answer options. If the answer options varied by their number of words, the correct answer must have all of the words in the answer option concept overlap with the returned web query. As a result, in this answer selection metric, longer answer options were less likely to successfully match the web data.

Query type *C* was also less successful in matching longer answer option concepts. The approach of querying an answer option with the text of the question stem yielded results not dissimilar to the other two methods. The correct answer was chosen based on the answer option and question stem query pair that returned the most hits. For average length questions and answer options, the results suggest that this is a viable approach. Perhaps this is because two-word answer options were correctly treated as a colocated word pair. When longer answer options were combined with longer questions, the results were not as good.

Some of these data analysis issues are exemplified by Example 6. The two highest scoring "Define: X" results in Example 6 supported "Ependymal Cells" and "Schwann Cells." Neither result is the correct answer but each shares the term "cells" with the

question stem. "Glia" is the more important word in the noun phrase "glia cells" and it appears as the root of the correct answer "Microglia." Unfortunately, this match was not caught because "microglia" was not separated into its constituent parts and there was no weighting of the word matches that would have supported a different, correct result.

Example 6

What type of **glia** *cells* engulfs and destroys *micro-organisms* and debris?

A. Astrocytes
B. **Microglia (correct answer)**
C. Ependymal *Cells*
D. Oligodendrocytes
E. Schwann *Cells*

The biology domain uses specific scientific terminology that often adheres to Latin and Greek-based word templates. "Micro" is a prefix that could be used with other similar biology prefixes to identify or decompose words for possible matching. Domain-specific spelling variants should also be incorporated. "Micro-organism," which appears as it was written by a student, could match "micro," "organism," "micro-organism," "micro organism," but most importantly, it should first match "microorganism." Similarly, the suffixes "-cytes" and "-cyto" mean "cells." Incorporating a table of biological terms and their variants would increase successful matching.

In addition, the best results from the query type experiments were higher than the best results of the Lucene system discussed in Section 4.2.2. This may be linked to the test set questions being procured from professionally authored question sets where in particular the questions contained well conceived and succinct question stems and answer options. A more detailed discussion of the results of the automated system and the methods for chosing the best answer is found in the next section. Most importantly, the analysis of this section justified my decision for going forward with "Define: X" as the primary methodology for use in the automated solution of MCQs.

4.2 Automated System Results

There are two themes that underly the work in this section. The first is how to validate the success of this automated question analysis system with human results. Answering questions where the answer is known is not always insightful. Answering hard

or differentiating questions is. Human results from sets of students answering sets of questions can provide just that information. The results of the experiments that incorporate human data, through Item Analysis and exam building, are presented in the following sections, 4.3 and 4.4.

The second theme is the validation and extension of the work presented in Section 4.1. Exploring the best way to answer an IDMCQ with web queries and bag-of-words overlap measures was a starting point to create a methodical approach to judging question difficulty. The results of Section 4.1 encourage deeper investigation into how best an automated system can match similar terms. The biological domain of the questions, their concept-focused format, and the use of the Web as an answer resource has many effects on the success of the automated system. The focus of Section 4.2.1 is finding the answer terms in data and picking the method to choose the best answer. Section 4.2.2 presents possible limitations to the bag-of-words approach used in this research.

The following section describes the results of running the PeerWise data through various versions of the automated query system, looks at the characteristics of the questions, and reflects on how to address improvements to the system.

4.2.1 ROUGE and Lucene Results and Discussion

There are two ways that the data was grouped for comparisons. The initial experiment in ROUGE used single queries with "Define: X" that returned 50 answer titles and snippets per answer option. That method was repeated for the Lucene experiments and is called *individual*. The second way of comparing the results was called *grouped*. In the grouped approach all 50 of the returned web titles and snippets are combined and indexed as a document.

To validate another feature of ROUGE which is based on bag-of-words overlap, the baseline Lucene index (also based on bag-of-words overlap) was implemented. In addition to bag-of-words, Lucene bigrams were used to attempt to find more colocated terms, like many of the two word concepts that make up the answer options. Finally, WordNet was introduced in an attempt to incorporate more possible lexical and semantically related word matches [64]. The result of this was to introduce additional semantically related words into the indexes used in the comparison.

Table 4.3 shows the different ways that the data in the titles and snippets was compared to the words in the answer options. There are three different methods for selecting the top answer and Table 4.3 shows that all of the answer choices have answer

options associated with the single highest score of term similarity except the ROUGE choice which is based on the top average score of the top 50 titles and snippets per answer option. A portion of the Lucene results may be found in Appendix I.

Table 4.3: Results using different data comparison methods based on the correct answer as selected by the top average, or "aggregate" similarity scores of all of the documents (web page titles and snippets) indexed with each answer option. A sign test showed that there were no significant differences between the approaches.

Course	1 Individual	2 Individual	1 Grouped	2 Grouped
ROUGE (BOW)	.4013	.4013	n/a	n/a
Lucene (BOW)	.3864	.4060	.4135	.4621
Lucene (BIGRAMS)	.4045	.4015	.4045	.4696
Lucene (WordNet)	.4198	.4242	.3969	.4469

Three different answer selection methods were implemented to test choosing an answer based on the comparison data. The first method was the single *top* scoring answer based on Lucene's document similarity measure. The second, called *aggregate* was based on the total average of all of the scores for all of the documents associated with each answer option. This method is most similar to the way that the top ROUGE scores were chosen. With ROUGE, all of the scores for each web query results were averaged and the answer option with the highest average was deemed the correct answer. The third answer selection method is called *hits* and that is based on counting the number of times the answer option occurs in the web results text. The results are shown in Table 4.4.

Table 4.4: Results using different answer selection methods. The comparison was based on Lucene bigrams. Observe the low percentage of "hits" indicating a dearth of matches returned from web retrieval.

Course	1 individual	2 individual	1 grouped	2 grouped
Top	.3969	.3863	.4045	.4696
Aggregate	.4045	.4015	.4045	.4696
Hits	.3129	.3257	.1832	.1287

The single highest similarity score was achieved with the Course 1 grouped data using the aggregate and top selection methods (there was a tie) on Lucene bigrams. This

similarity score corresponded with four answer option questions. The single highest similarity score for one complete set of course questions (all numbers of answer options) was the Course 2 grouped data, using the top selection method with Lucene bigrams.

The best automated system results are more than three times the expected results from randomly choosing the answers in a MCQ with five answer options. Still, they provide ample room for future improvement based on a deeper analysis of the questions that worked and those that were incorrectly answered with the automated system.

Table 4.5: Questions in Course 1 and Course 2 grouped by the percentage overlap of text shared between the question stem and the definitions of the answer options.

Match results under	2.5%	5%	10%	15%	20%
Number correct	5	5	14	2	0
Average question difficulty	67.47	70.22	67.66	57.65	n/a
Number with + discrim. power/average	3/.023	5/.235	10/.188	2/.282	n/a
Number incorrect	5	10	18	3	1
Average question difficulty	68.48	62.54	61.85	69.11	20.3
Number with + discrim. power/average	2/-.024	7/.212	12/1.65	2/.94	1/.071

An overview of the matching results for both Course 1 and Course 2 are shown in Table 4.5. The questions are broken down within each threshold by whether or not the automated query system returned the correct answer. There is also the average question difficulty and the number of questions with positive Discrimination Power and the average Discrimination Power for the questions. The discrimination measures are based on Mitkov et al. [65] and described in greater detail in Section 4.3.

In the test set using "Define: X," 40.13% of the questions were answered correctly. In instances where there were ties, and one of the tied answers was the correct one, I gave fractional points based on how many answer options the tied answer tied with (.5, .3, or .6). There were no more than three answer options tied. None of the questions failed to match at least one of the answer options. When ties were considered, 42.64% of the questions were answered correctly. Random guessing would answer only 20% correctly.

Splitting the questions into groups based on how much information their answer was based on reveals how a majority of the successful matches depend on very little information. The most matching question stem to answer option definition had only 1

out of 5 words in common. The average question difficulty is consistently in the 50 to 70% range except for the match results under 20% which has a 20.3% average because they are based on one question. Example 11 is also the most difficult question in the two PeerWise sets and discussed in more detail in Section 4.3.3. As a reminder, these matches are not based on common stop words but rather bag-of-words overlap of the content words in the definitions.

Analysis of the student exam results suggests a method for classifying the output of the automated system that parallels the human results. This supports my underlying belief that the closeness of answer options to the question stem and each other is reflected in student results. When a student is choosing an answer to an IDQ, he or she is in some manner choosing the closest concept, process, or term that sufficiently fulfills the description given in the question stem. Based on this perception, a difficult question is one where the answer options are not only closely related terms but also meaningfully linked to the question stem. In other words, my research compared how students selected answer options with how the automated system found shared terms in the definitions of the answer options and the original question stem. In order to learn how these experiments could be improved based on the results, I look at the following factors: how the questions are compared, a group of difficult questions, a group of discriminating questions, a group of questions where the match could be improved, and a group of questions that are outside of the current question parameters. An analysis of question difficulty, discriminating power, and distractor usefulness is discussed further in Section 4.3.

4.2.2 Some Limitations on the Bag-of-Words Technology

I have performed more than 20 experiments using bag-of-words technology in the introduction to Biology domain (not all of them are included in this thesis although many are located in Appendix I). My research used a variety of selection methods to choose the best answer, two approaches to group the answer data (*individual* and *grouped*) and two different bag-of-words based software systems. WordNet was also introduced in an effort to increase the terms that would be possible matches. Nonetheless, the results of the automated question analysis system are roughly 40-42%.

In the test set experiments, the average success of the three query types was 54%. What is the basis for the discrepency between those results and the lower ones on the PeerWise data? As mentioned in Section 4.1, the test set data were created by

professional test makers and may have had more articulate and succinct question stems than the PeerWise data. More on this discussion can be found in Chapter 5.2.2.

If the automated system results were better, perhaps in the 60% and higher range, I would feel more comfortable lauding the merits of running new data through this system. I contend that the rigor of the answer matching, answer selection, and potential answer data collection aspects of the automated system are sound, so the answers must not be in the data that was collected to answer these questions. In my opinion, there is a limit to how much information may be gleaned from the Web especially in a specific domain such as introductory biology. In my opinion, to answer more of these questions correctly, structured data from the biology domain is needed to augment the pool of data where the answers may be found. Early Question Answering systems that used structured data were introduced in Section 2.2.1. The current state of the art may be the work by Chaudhri et al. on developing an "intelligent" texbook that could answer questions similar to those in the PeerWise question stems by searching the structured knowledge base of a digital biology textbook [29].

Another way of improving these results might be to incorporate human-in-the-loop judgments in the proposed automated question analysis system. Further, access to the proprietary databases of for profit testing companies could introduce measurable improvements. Finally, the limitations of the automated question analysis system have an important effect on the primary goal of this research, automatically identifying question difficulty, discrimination power, and distractor usefulness. As mentioned in the previous paragraph, there may be other ways to collect data that contains the answers to these questions from a biology-specific knowledge base, even one such as Wikipedia's biology resources. The limits of the automated system affect the potential links to human performance data that is discussed in the following two sections.

4.3 Human Results from Item Analysis

Sections 4.3 and 4.4 are based on research begun in [96] and extended more recently in [75] [1]. That research in turn is based on the early exploration of Test Item Analysis by Davis and Gronlund [76] [26] which was incorporated into question generation and analysis work by Mitkov et al. [65]. Human performance data provides valuable information about question difficulty, question discriminating power, and the usefulness of distractors. In this work, PeerWise data provided the human judgments that allow Item Analysis to model what makes a question difficult or discriminating. The underlying

methodologies used in the following sections were presented in Sections 2.2 and 3.3.

4.3.1 Results and Discussion

Mitkov et al. [65] introduces distractor classes, which are a way of grouping how well an answer option "distracted" lower-performing students as opposed to their higher performing peers. The four distractor classes are "good" or useful distractors, "poor" distractors, not useful distractors, and distractors that confer no discriminating power. These terms were introduced in Section 3.1, but because they are slighty counter-intuitive, they are summarized here.

Good distractors are negative numbers indicating that they attracted more low-scoring students than high-scoring ones. Poor distractors are positive numbers because they are chosen by more high-performing students than low-performing ones. The goal of exams is to differentiate performance groups of students and distract less-prepared students. Distractors that had equal numbers of low and high-performing students choose them are called non-discriminating distractors because they failed to separate the students into performance cohorts. Distractors that were not chosen at all are considered not useful.

My hypothesis for generating a model of question difficulty is focused on how many distractors with negative (good) values there are for a given question. One way that MCQs can be hard is when there are several strong answer options. When multiple answer options have positive numbers, those are answer options with strong distraction power. In this case, good students act like their lower-performing peers conventionally do and spread their answer choice among several answer options. Difficulty can be measured by the total number of students who answered a question correctly divided by the number of students who took the exam. Difficulty can also be measured by how many good distractors a question has.

Discrimination, on the other hand, reflects how well a question sorted students into the three performance cohorts. The concept of performance cohorts was introduced in Section 3.4. An ideal exam quickly and correctly places students into their deserved group. From an educator's perspective, the more discriminating a question, the better it is. By performing Item Analysis and examining the discriminating power of individual questions, my research seeks to find a point in an exam where a student has answered enough questions to reveal what performance group he or she is in. In other words, since the process of discrimination is to split the students into three performance

groups, how many questions does a student have to answer for his or her performance group to be identified? This approach aims to minimize the number of questions that students need to answer in exams. MCQ-based exams fundamentally seek to measure comprehension. Once a student has answered enough questions, he or she is grouped into an achievement cohort. Ideally, automating the creation of highly discriminating questions would vastly reduce the number of questions a student would need to answer before his or her performance group is identified.

Appendices G and H contain examples of questions processed by Item Analysis using Mitkov et al.'s distractor methodology [65]. To indicate the distractor classes in the "UFN" (usefulness) column in the appendices, "X" represents the correct answer, "N" indicates a distractor was not useful, and the rest of the distractor classes are shown with the polarity of the difference of the high-performing students who chose an answer option minus the low-performing students who chose the same option.

Course Number	**Course 1**	**Course 2**
Average Question Difficulty	0.5801	0.6725
Instances of Too Easy (greater than .85)	0	3
Instances of Too Difficult (less than .15)	0	0
Average Item Discrimination Power	0.1964	0.1366
Instances of Negative Discrimination Power	2	5
Instances of Poor Distractors	67	41
Instances of Not Useful Distractors	22	33
Instance of Useful Distractors	40	45
Instance of No Discriminating Power (0)	9	6
TOTAL Number of Distractors	176	159
TOTAL Number of Questions	37	33

Figure 4.1: Results of performing Item Analysis on two exam-like sets from PeerWise Courses 1 and 2. The ideal question difficulty is .5 and the maximum positive discriminating power is 1.0. Maximum positive discriminating power occurs when all of the students in the high performing group answer a question correctly and none of the students in the low performing group do.

Figure 4.1 presents examples of processing two exam-like sets from the PeerWise data. The average discriminating power in both courses is lower than might be desired. Ideally, discriminating power is as close to 1 as possible which means that the questions successfully differentiate students into performance groups. A negative value for discriminating power, as evident in 7 of the 70 questions in this experiment, means that students who in general performed less well than their peers correctly answered this question more often than the better performing students. A comparison of the distractor classes also shows that poor and not useful distractors combined occur twice as often as useful distractors. In addition, 4.5% of the distractors had no discriminating value.

The Item Analysis results shown in Figure 4.1 are from two exams with question difficulty greater than the ideal of .5, where 50% of students get an answer correct. Nonetheless, there were only 3 questions in the sets considered "too easy", where the threshold for too easy is greater than or equal to 85% of the students getting it correct. Similarly, "too difficult" questions are answered correctly by fewer than 15% of the students who try them. Similar results were reported on by Mitkov et al. which was used as a basis for these comparisons [65] [64]. Because both the problem domains (Biology and Linguistics) and sets of students are different, it does not make sense to look for significant differences within the numbers themselves. Nonetheless, this comparison does suggest trends that help identify a good question in an exam.

Gathering student-question data sets, creating exams out of them, and performing Item Analysis have all focused on discovering what makes a question discriminating and what makes it difficult. Internal Item Analysis is a tool used to examine how individual questions are difficult and discriminating in the context of an exam. An extension of Internal Item Analysis that looks at human data that is not available in this work is External Item Analysis. External Item Analysis examines overall student grades and rank across several exams for correlations between how a student performed on an exam and how well he or she did in the class as a whole. One facet of External Item Analysis is that it can be used to flag students who are high performing via other measures, but test poorly [26].

4.3.2 Filtering for Discriminating Questions

Adjacency matrix-based exam creation is simply a data preprocessing step that provides potential evaluation materials in a faster and cheaper manner than by manual

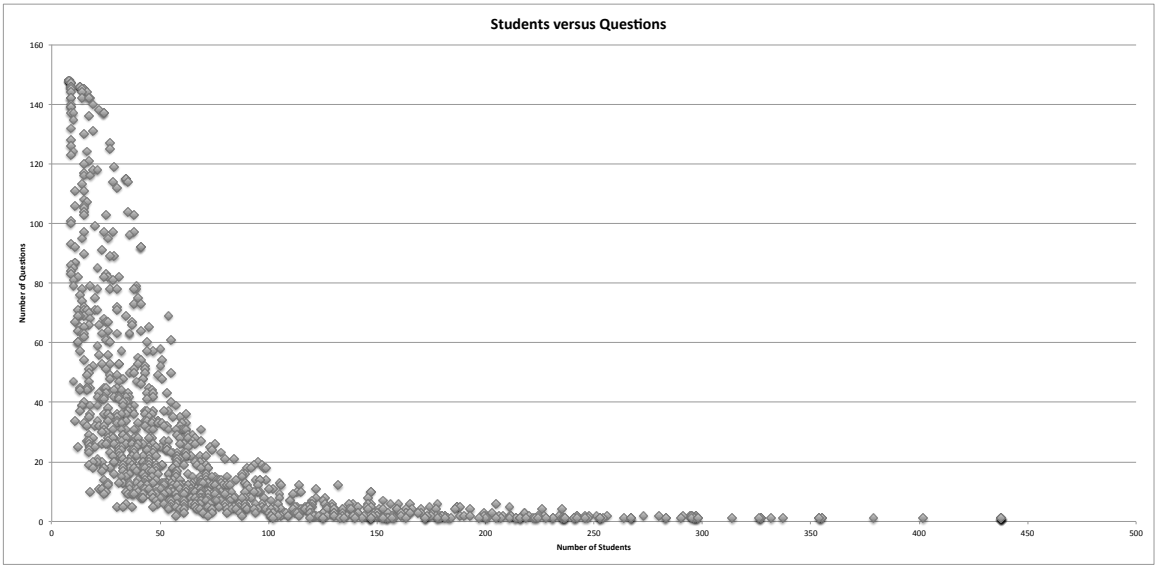

Figure 4.2: A set of results for Course 1, which shows a set of potential exams. Each point represents a unique new exam for a given number of students and a given number of questions. The x-axis represents the number of students and the y-axis the number of questions.

methods. There is nothing in this approach that identifies which questions are discriminating *a priori* because a complete exam is first needed to make that measure. The steps to filtering for discriminating questions are:

- Perform Item Analysis to attain the question difficulty, which is shown in column 6 of Figure 2.6.
- Examine how many answer distractors fit into each distractor class as described in Section 3.3.
- Count all of the distractor class instances and average the question difficulty and item discriminating power as presented in Figure 4.1.
- Rank the questions based on the highest to lowest question discriminating power.

Once the exams have been created, grade them and sort the students into cohorts based on their performance. Next, Item Analysis shows how many students correctly answered the question compared to the number that tried, as shown in column six of Figure 2.6. This column corresponds with question difficulty, which is in the example in Figure 2.6 is 35%. The lower the percentage, the more difficult the question. It is not simply the most difficult questions that do not discriminate, since questions that are

too easy also fail to effectively group students based on their performance level. The question sets were previously reviewed for questions that were too easy or too difficult. There were no questions in the data set that all of the students answered correctly. Similarly, there were no questions that all of the students answered incorrectly. If there had been nondiscriminating questions that all of the students answered in the same way, those questions would have been eliminated from the set.

Overall, the PeerWise Course 1 and Course 2 students answered about one third of all the questions incorrectly. A naïve approach would force a threshold of question difficulty onto the exams so that only sufficiently hard questions would be used to support the automated system. In one experiment based on thresholds presented in Mitkov [64], 85% difficulty was the threshold for question inclusion, and only questions below 85% were added when creating the exam. Again, the lower the percentage the more difficult the question. This approach reduced the upper bound on the question-student exam size from 148 questions answered by the same 9 students to 144 questions answered by the 9 students. While 85% difficulty is considered *too easy*, 15% difficulty is considered *too hard* and both of these thresholds are included in the Item Analysis measurement methodology of Mitkov et al. [65]. Figure 4.1 shows counts of *too easy* and *too difficult* questions in the Course 1 and Course 2 "exam" sets.

All of the steps in exam building up to this point have been focused on testing the individual questions in an exam for their difficulty. In the two examples from PeerWise presented in Figure 4.3, the first question, 34905, had a difficulty of 63% and the second question, 31761, had a difficulty of 42%. Question 31761 is therefore the more difficult question.

Discovering the most discriminating questions in an exam is another goal of my research project. Discriminating questions most effectively sort students into their relative cohort, and are the most valuable questions in exams because they provide the most information about a student's level of comprehension. Question 34905 is an example of a more discriminating question (discriminating power .282) than question 31761. Answer C, the correct answer, was chosen by a large majority of the high-performing students as well as the students in all other groups, which initially leads one to believe the question to be non discriminating, but when the distractors are taken into consideration, all four of them attracted more lower performing students than high performing ones. In other words, this question managed to sort students into performance groups because the poorly performing students performed similarly, and chose distractors different from the high performing students.

Question 31761

A sperm cell about to undergo meiosis II is called a?

A. Spermatagonia
B. Secondary spermatocyte *correct answer*
C. Oogonia
D. Primary spermatocyte
E. Spermatid

Letter	High	Middle	Low	Total	UFN	
A	13	7	4	24	9	
B	17	36	14	67	X	Correct
C	1	3	2	6	-1	
D	7	12	13	32	-6	
E	4	10	3	17	1	
OMIT	0	5	7	12		
TOTAL	42	73	43	158		

Discriminating power: .070
Omission rate: 0.076
Question Difficulty: .4241

Question 34905

Which substance exists in the crystalline micro-fibrillar phase?

A. Pectin
B. Extensin
C. Cellulose *correct answer*
D. Lignin
E. Hemi-cellulose

Letter	High	Middle	Low	Total	UFN	
A	3	4	6	13	-3	
B	0	6	3	9	-3	
C	33	45	21	99	X	Correct
D	3	8	4	15	-1	
E	3	8	5	16	-2	
OMIT	0	2	4	6		
TOTAL	42	73	43	158		

Discriminating power: .282
Omission rate: 0.038
Question Difficulty: .6266

Figure 4.3: Example questions 31761 and 34905 from the PeerWise data followed by tables supporting Item Analysis computations.

Questions 34905 and 31761 are shown with their results from the Item Analysis that was presented earlier in Figure 2.6. The Item Discriminating Power, omission rate, and question difficulty are calculated below the questions and reveal the value of the question to the exam. For example, in question 34905 the most omissions (4) came from the low performing group.

Question 34905 was answered 403 times and it was answered correctly 114 times in the unfiltered, sparse data set. In this exam based on 158 students (the top 15% most correlated students), the three cohorts are shown on the row titled "Total." The 27%-46%-27% cohort split for high, middle and low-performing students translates in this exam size to 42 high, 73 average, and 43 poorly achieving students.

In Figure 4.3 Question 31761 is an example of a more difficult, but less discriminating question than question 34905. Question 31761 has 2 poor distractors (positive value) and 2 good distractors (negative value) in the UFN or usefulness column. UFN shows the difference between what answer options the high and low-scoring students chose and produces a zero, a negative, or a positive number. Zero means a question has no discriminating power, if no students chose it. If equal numbers of high and low performing students chose it, it is a non discriminating question. A positive number means that lower-scoring students were more attracted to this option than their higher-performing peers. Negative numbers in the UFN column are the marker of a good distractor. A more difficult question usually has more than one good or negative

valued distractor.

Appendices G and H contain information like that shown in Figure 4.3 for 70 questions used in the exam building that is the focus of Section 4.4. In addition to Item Analysis data there is also human ratings on how difficult and how good the questions are. These ratings are part of an extensive amount of information, enabled by the PeerWise environment, collected during the question authoring and answering. Only a small fraction of this information was directly relevant to discovering what aspects of answer options makes a question difficult and further what characteristics of these questions make them discriminating. As mentioned in Section 3.3.2, the relevant question materials consisted of the unique question ID, the unique student ID, the average rating for each question (0 to 5), the average difficulty for each question (1 to 3), the total number of responses, the total number of ratings, the correct answer, the number of answer options, the text of the question, and the text of the answer options. The relevant answer materials are what questions each students took, how they answered them, and how they rated them for difficulty and quality. Again, while interesting, the human ratings were not used in this current research. They are further discussed in Section 5.2.

4.3.3 Difficult and Discriminating Questions: Further Examples

As noted in Section 4.3.1, in addition to looking at how questions are compared, I also look at a group of difficult questions. Example 10 appears in Appendix D as question 41354:

Example 10

What hormone has a negative influence on growth?

A. Thyroid Hormone
B. Insulin
C. **Cortisol (correct answer)**
D. Testosterone/Estrogen
E. ACTH

Only 1 out of 3 students answered Example 10 correctly. This question is difficult in a number of different ways. The question stem includes the phrase "negative influence," which does not violate the "no negation" question rule because it does not correspond to an absence like many negative questions do. In Example 10, the word "negative" is not used to describe the absence of a property like in negation questions. In the

case of "negative influence," "negative" is being used as a modifier of the noun "influence." This distinction is important because this question is seeking a property, not the absence of a property.

Nonetheless, for the automated query-matching metric to work, this noun phrase needs to occur in one of the definitions of the answer options. Unfortunately, it does not occur and the distractor "Thyroid Hormone" is the top result from the automated query pipeline.

Both students and the automated system failed to answer this question correctly because it is a difficult question. From the system's point of view, the key terms "hormone," "negative," "influence," and "growth" are being matched to the results of "Define: X" on the answer options. This is a short question with only four key terms. Another complexity is that the correct answer, "Cortisol," is commonly referred to as "hydrocortisone," which would require either the identification of the common term "cortiso" or a biological terminology thesaurus to associate a connection between the terms.

From the student's point of view, Example 10 is hard for another reason. The use of "Cortisol" is confusing because "hydrocortisone" is both a more common and a more formal variant of the term. Beyond the specific terminology, all of the other terms have close meanings. They are all hormones that have some relationship with metabolism, growth, and cell stimulation. Even adrenocorticotropic hormone (ACTH) is linked to the opposite of what the question seeks: An increase in the growth of body extremities (Cushing's syndrome). This question is hard because the answer options are conceptually close. More information can be found on Example 10 in Appendix D, where it appears as question 41354.

Even more difficult than Example 10 is Example 11, which is question 41332 from Appendix H.

Example 11

Which primary hormone is responsible for growth in a one-year-old child?

A. Growth Hormone
B. Cortisol
C. Insulin
D. **Thyroid Hormone (correct answer)**
E. IGF-1

While "Growth Hormone" is not the correct answer, the words "growth" and "hormone" do occur in the question. This question was unsuccessfully answered by both

the automated query pipeline, and the students who attempted it. With 1 out of 5 students choosing the correct answer, this question is the most difficult of the 63 questions in the unfiltered PeerWise question sets.

In contrast to the features of difficult questions, Example 12 which is question 35681 in Appendix G, is a discriminating one:

Example 12

A defect in the chloride ion transporter channel is responsible for what disease?

A. Type II Albinism
B. **Cystic Fibrosis (correct answer)**
C. Wilson's Disease
D. Epilepsy
E. Neurofibromatosis

In Example 12, the Discriminating Power is .188, which means that in the case of every answer option, more high-performing students chose that answer than did lower-performing students. In Example 12, there were considerably more omissions in the lowest performance group. Thus, this question is probably not as discriminating as it seems, since most of the people in the low performance group did not actually attempt the question. Because high omission rates negatively affect Discriminating Power, there could be improvement in the way that question discrimination is currently measured.

In order to achieve the goal of obtaining question discrimination it is important to reduce or refine non discriminating questions in exams so that tests are comprised of fewer, better questions. The discrimination analysis shows, based on the methods used to build exams presented in Section 4.4, that these questions had high omission rates. A discriminating question is one where there is not an obvious answer but two or more good, challenging answer options. Thus, a discriminating question could be one with a Discrimination Power approaching 1. This method can be improved with more subtle measures that weight questions with high omission rates.

4.4 Exam-Building Results

As mentioned at the beginning of Section 4.3, the results and discussion in Section 4.4 follow the arguments published on building exams [96] and detail the difficult

questions and high performing students in exams [75] [1]. As mentioned in sections 2.4 and 3.3, Item Analysis uses human data to produce measures of question difficulty, discrimination, and distractor usefulness. In the following sections the results of the two methods used to build exams are described, the matrix (or clique-based), and weighting-based approaches. Exam building is an important step that turns a set of questions into exams ready for further processing with Item Analysis.

Section 4.4 presents the optimal parameters for applying the clique-based methodology and describes the results. This section also shows how the question weighting-based methodology provides lower quality results when compared to the matrix-based approach. This section ends with observations about analyzing cohort movement while creating useful exams.

4.4.1 Matrix-Based Results

Section 3.4 introduced the role of cohorts in building exams, the matrix-based method of exam building, and the weighting-based approach. Figure 4.4 shows how student cohort movement develops as students answer each additional question. The 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. The low-achieving student whose path is closest to the x-axis and ends with 3 correct questions answered, is quickly separated from the performance pack. The middle-performing student who finishes with 7 correct answers is in the same cohort as the top student, shown farthest from the x-axis, until the forth question is answered.

In Figure 4.4, if new, additional students were clustered in the regions around the students (shown on the right hand side of the figure) they would be in the same performance group. Thus, by adding all of the students taking an exam, these additional students would reside along the performance range of high-, middle-, and low-achieving students. Each performance group is distinctly differentiated in Figure 4.4 and this representation of how students perform may suggest re-ordering questions to expose the cohorts more quickly. Cohort movement occurs when a student is in one performance group (like with the top and middle performing students in Figure 4.4 until question 4), and then their question responses sort them into another cohort.

During my research I discovered that student movement between performance cohorts is quite high, around 30% when there are very few questions in the exam. Because I desired stability in performance cohort movement, I used 15-30% of both cor-

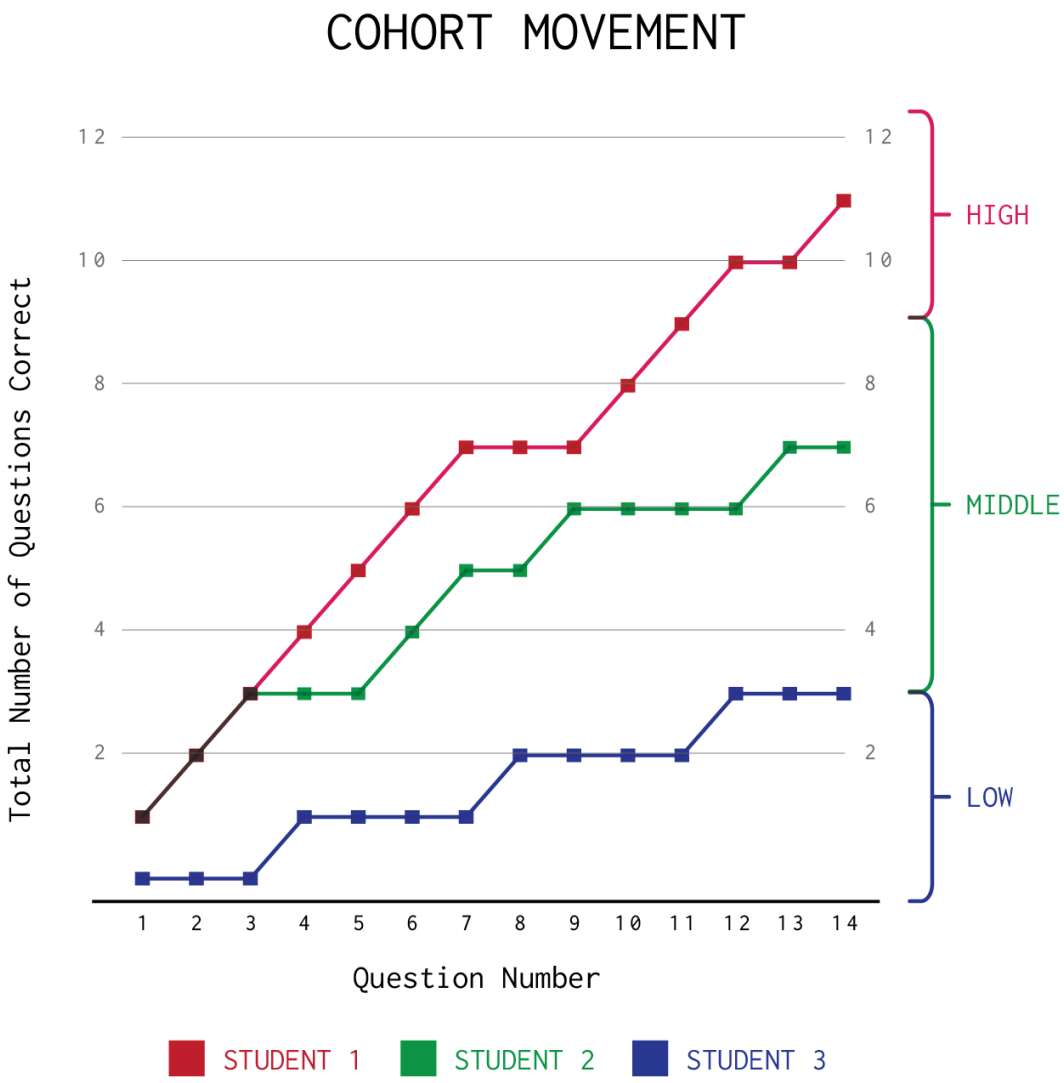

Figure 4.4: Movement within the cohort. 3 students move into different cohorts based on the number of questions that they have answered correctly in comparison to other students.

related 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. There is a more extensive discussion of my approach in Appendix D, and Figure D.4 demonstrates the results of balancing the most correlated students and questions. After creating the exams from the most correlated 15% of the students and the most correlated 25% of the questions, Item Analysis was used to find the most difficult and most discriminating questions within the exams.

4.4.2 Question Weighting Methodology

As shown in Figure 4.5 on the left, 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, it was discovered 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 [26] 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 very low and very high weights would not have much discriminating power, but when this method was applied, cohort movement was unacceptably high. A similar reasoning follows that very easy and very difficult questions fail to effectively differentiate students into cohorts because so many or so few students get them correct. It is also important to note that one of the benefits of the question weighting approach is that it is extensible in situations with very sparse data. The weighting approach does not compare favorably with the clique-based method for exam building in this situation but the clique-based approach is less effective in very sparse data environments.

In the question weighting method, students who answered fewer than 3 questions were omitted from the analysis. Their answers were not considered in the question

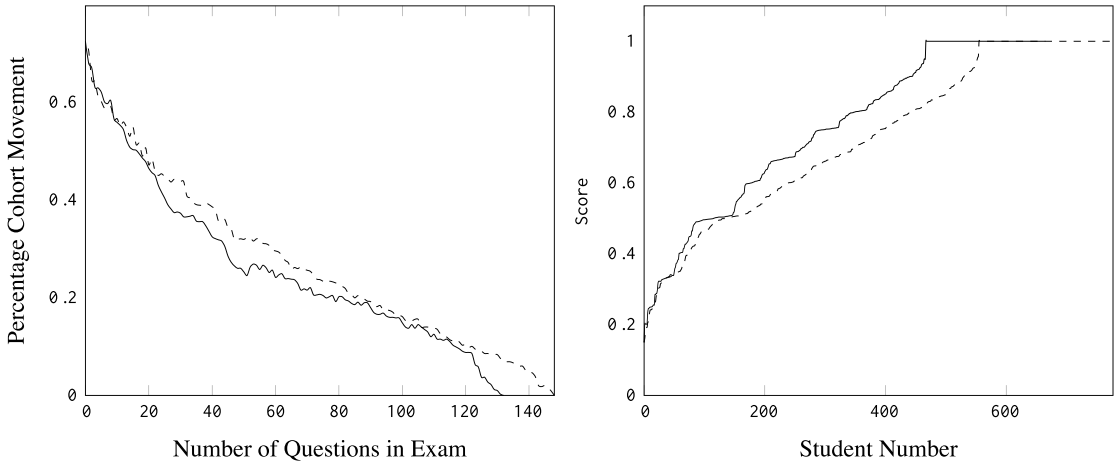

Figure 4.5: The graph on the left shows cohort movement for Course 1 and Course 2 where Course 1 results are indicated by the dashed line. Course 1 starts with 148 questions and Course 2 starts with 132 questions. When the most discriminating 26 and 20 questions remain, cohort movement is 44% and 46%, respectively. The graph on the right shows student performance for Courses 1 and 2 when 26 and 20 questions remain, respectively. Course 1 is indicated by the dashed line. In this data set, many of the students answered all questions correctly, shown by an ending plateau.

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 Figure 4.5 on the right, the performance results show that in Course 1, over 50% of the students scored 75% or better. In Course 2, 44% of the students scored 75% or better. These results indicate that better students answered more questions and these results will be further discussed in Section 5.2.

4.4.3 Steps Towards Creating the Ideal Exam

After the MCQ analysis of Section 4.3.2, we can now ask which and how many questions are needed to create a quality exam. I presented two methodologies for creating quality exams. Both approaches filter out the least discriminating questions in an exam in an attempt to improve test efficiency. The first approach analyzed the best balance of students and questions based on creating a more dense matrix of those students and questions. The second approach initially analyzed the questions' difficulty to find the best new exam set with the most discriminating questions.

Meanwhile, there must be a sufficient number of questions in an exam to effectively

divide, using Item Analysis, the performance groups into three cohorts. Ideally, there are far more questions and students included than this minimum. Figure 4.5 presents all of the possible exams that could be created from Course 1 data. Figure 4.5 shows what the Course 1 data looks like before using clique-based or weighting-based methods of creating a more complete exam.

Course 1 began with a sparse matrix of 1055 students and 148 questions. Using the adjacency matrix approach, I created a new exam composed of 158 students and 37 questions. Once I discovered a sufficient new exam size I performed Item Analysis and measured the discriminating power of each question. The filtering is based on a question having a positive discriminating power as described in Section 4.3. The maximal discriminating power is 1 so questions that have discriminating power scores that approach 1 are ideal. After correcting for item discriminating power, the new exam size was 26 questions answered by 158 students.

Course 2 began as a sparse matrix of 887 students and 132 questions and using the matrix approach became a new exam composed of 158 students and 37 questions. After correcting for item discriminating power, the new exam size was 20 questions answered by 133 students. The new exams for Course 1 and Course 2 have characteristics that are shown in Figure 4.6.

Method:	Clique	Clique	Weighted	Weighted
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

Figure 4.6: Characteristics of the new exam including the numbers of questions and students before filtering for discriminating questions and the numbers after. Also, the movement of students from different cohorts is presented with 18 students moving in total. The weighted method eliminated students who answered fewer than 3 questions.

In Course 1, the cohort movement in this new exam was low at 11%. In this case, 18 students moved from one performance cohort to another based on how they answered the first set of questions compared to the second set. In course 2, the cohort movement was 20%. In both courses, the first set of questions included the top 15% most correlated students and the top 25% most correlated questions. In the second set, the questions deemed the least discriminating, based on the techniques described in Section 4.3, were filtered out.

In both courses, the allowed answer omission rate in the exam was 100% using the clique-based approach. This permits students who did not answer the question, but were a member of the top 15% most correlated students to participate in this exam. This rate may vary in future experiments. One problem with permitting omissions was that the members of the lowest performance cohort in many questions were good students who just did not happen to answer that particular question. Omissions are not a factor in the weighting-based approach because it is a given that many of the students did not answer questions in common. Note that the cohort movement only occurs in this instance from the cohorts next to one another, such as the middle to high. No students move from the low to the high performance group or vice versa. The complete Item Analysis results for both sets are presented in Appendices C and D.

Because I wanted stability in performance cohort movement, I used a percentage of both correlated students and questions that was between 15 and 30%. 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 I created the exams from the most correlated 15% of the students and the most correlated 25% of the questions, I used Item Analysis to find the most difficult and most discriminating questions within the exams.

In Course 1, I began with a sparse matrix of 1055 students and 148 questions and using the adjacency matrix approach created a new exam composed of 158 students and 31 questions. Once I discovered a sufficient new exam size, I performed Item Analysis and measured how discriminating each question was. The filtering was based on a question having a net positive Discriminating Power as described in Section 4.3.2. After correcting for Item Discriminating Power, the new exam size was 26 questions answered by 158 students.

The cohort movement in this new exam was low at 11%. In this case, 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. The first set of questions included

the top 15% most correlated students and the top 25% most correlated questions. In the second set, the same percentages of most correlated students and most correlated questions were used, 15% and 25% respectively.

The second set (Course 2) started with 887 students and 132 questions. Using the correlation matrix approach, the most correlated students and questions filtered the original exam size to a new exam of 133 students answering the same 31 questions. Then, Item Analysis was performed on this new exam. Similar to the first set, in the second set the questions deemed to be the least discriminating, based on the techniques described in Section 4.3.2, were filtered out of the exam. This filtering reduced the new exam further to 133 students answering 20 questions. The bin movement for Course 2 was 20%.

The results of Item Analysis are described in Section 4.3.1 and an example of both a discriminating and a nondiscriminating question is shown in Figure 4.3. The full Item Analysis results of Set1 are located in Appendix C. Performing Item Analysis reduced a possible exam set of 1055 students and 148 questions to 158 students and 31 questions. When the 33 questions were filtered to leave only the discriminating questions as described in Chapter 4, the final exam size was 24 questions. Performing Item Analysis reduced a possible exam set of 887 students and 132 questions to 133 students and 33 questions. When the 32 questions were filtered to leave only the discriminating questions, the final exam size was 20 questions.

Of the 63 questions in Set1 and Set2, 26 of them were answered correctly by the automated system. This 41.27% answer rate corresponds to the results from the experimental set used for building the system and is described in more detail in Chapter 3 and Section 5.1. When the question sets are filtered to only include questions with a positive discriminating power as described in Chapter 4, 20 out of 44 questions were answered correctly, or 45.55%.

Chapter 5 summarizes my research and introduces questions for future work.

Chapter 5

Summary and Future Work

The final chapter summarizes the research approaches taken in this dissertation and presents the conclusions reached. Finally, future work is suggested that could extend these results.

5.1 My Approach and Conclusions

Making exams is expensive and time-consuming. Methods that reduce the time and resources needed to create new, high-quality MCQs have value for both educators and students. Educators benefit by being able to tune their questions to the appropriate levels of discrimination and difficulty with less revision. Students benefit by spending less time answering questions since fewer questions are needed in exams when the questions are more discriminating. Two ways in which the "goodness" of questions are measured is by their difficulty and their discriminating power. This dissertation introduced the task of automatically assessing the difficulty and discriminating power of Multiple Choice Inverse Definition Questions.

First, I identified the role of Inverse Definition Questions and Multiple Choice Questions in standardized exams. While recent Question Generation research has focused on the generation of both question topics and answer options, I sought to best assess the difficulty of questions based on their answer options. Thus, I designed an automated question answering and difficulty pipeline to gauge the closeness of the definitions of the answer options to the question stem.

Then I compiled a corpus of Inverse Definition MCQs and answer materials to use as data to test the automated answering pipeline. These questions are from New York State Regents, SAT, CLEF, and Advanced Placement Exam websites and study prepa-

ration books. Initially, I narrowed down the focus to cover "also-known-as" and "is-called" inverse definition questions and close variants based on noun and verb patterns, in the biology and psychology domain. These "also-known-as" and "is-called" questions comprise a majority of the general inverse definition questions and the largest uniform set of all of the MCQs in the corpus. The test set data was used to judge the initial effectiveness of the automated answering system. Over 40% of the test set questions were answered correctly by the preliminary automated system. Similar results were found when the PeerWise data sets were also run through the preliminary system.

The original test set, while instrumental in the development of the automated answering pipeline, did not have any associated information on how difficult the questions were or how students actually performed on them. Having this additional information allows empirical analysis of how the automated pipeline performed as compared to a student class. Without human judgments on the questions I would not be able to tell if the questions my automated system answered were relatively difficult or easy. Seeking sets of traditional exam questions with student performance information, I discovered the PeerWise question bank. PeerWise contains existing MCQs collected in student-created question banks. The PeerWise data is sets of questions authored by students and answered by other students. I used two of these sets (*Set1* and *Set2*) from an introductory college-level biology course.

To analyze student performance, I needed to process the human results for their difficulty and discrimination power which I did in two ways. The first was to weight all of the students who answered 3 or more questions by how they performed on all of questions. The questions themselves were also weighted by how difficult they were based on how everyone who tried each one performed. The questions were ranked in terms of difficulty as were the students who were ranked and split into three performance cohorts. The second is the adjaceny-based method where I seek highly-correlating students and questions, or a set of questions answered by the same set of students.

In Chapter 3, I presented the weighting-based and the adjacency-based methods of creating exams so that I could use these human judgments to inform the results of my automated question answering and difficulty measurement system. To turn sparse sets of questions answered by some students into a highly correlated set of students and questions answered by those students, I employed adjacency matrix-based methods that are normally used to find maximal cliques in bipartite graphs. Meaningful student and question performance data is dependent on analyzing exams that contain

the largest number of questions answered by the largest group of the same students. A set of questions answered by the same set of students is often referred to as an "exam." Thus, the "best" new exams attempt to maximize for both students and questions. The adjacency matrix-based approach allows approximating a solution to this NP-hard problem and is extensible to other data sets. In addition, the adjacency matrix-based method measures how many questions are needed in the exam for the students to stop moving between performance groups. Distinct membership in a performance cohort indicates strengthened significance of student performance data based on the new exams.

In the educational psychology literature, one way of evaluating question difficulty and discriminating power is with Item Analysis. Both the weighting-based and the adjacency-based method use Item Analysis to analyze the students' performance as a group and return individual question difficulty and discrimination information. Item Analysis looks at how every student in a class performed on an exam and splits the class into three cohorts: the highest-, middle-, and lowest-performing students. Measuring question difficulty is based on the total number of students who choose the correct answer as compared to all of the students taking the exam. Then, the answer option choices of the top-performing students are compared to those choices of the lowest performing students in order to measure question distractor discriminating power. A discriminating question is one that is answered correctly by more high-performing students than lower-performing students. These questions are also described as having positive discriminating power.

After running Item Analysis on these new, crowdsourced exams from the PeerWise data, I measured both question difficulty and discriminating power. Filtering the crowdsourced exam questions for only those with positive Discriminating Power resulted in an exam constructed of the best, most discriminating questions. Finally, I ran the newly gathered exam questions through my automated question answering system as described in Chapter 3. I sought how these models of exemplary (difficult and discriminating) questions corresponded to the automated similarity metric results of the system.

Chapter 4 describes the results presented by three sets of experiments. The first set of experiments were the preliminary ones used to inform the design of my automated system. In the second set of experiments, I built an automated system that used both bag-of-word and latent semantic analysis-based methods of comparing possible answers to MCQs. The third set of experiments was based on Item Analysis and two

approaches for building exams. The two exam building methods are matrix-based and weighting-based. I refer to the third set of results as the human results because they are the benchmark used for question difficulty and discrimination.

I have presented MCQs in my research that have four and five answer options [A-D or A-E]. One assumption is that a student will chose one of the listed answer options but in real-life test taking that is sometimes not the case. Exams are constrained by time and in other cases by penalties for incorrect answers which encourage students not to guess on questions that they are unsure about. Both of these situations produce exams with skipped questions because students are moving on to questions that they feel more confident about answering. If one half of an exam contains four answer-option questions and the other half contrains five answer-option questions, the exam could be considered to contain five answer-option questions and six answer-option questions if an omission is counted as an option. Using an omission as an option reduces what would be chosen at random from 22.5% to 18.5%. Thus, my automated system answers questions almost three-times as well as random selection.

Software that emulates successful test designers and motivated test takers is valuable. I created new exams from crowdsourced online data in an effort to compare the output of automated experiments with that of a human gold standard. This approach successfully identifies two aspects of "good" questions based on human performance and notes that the automated results positively correlated with that human performance.

5.2 Future Work

There is always more to be done. Analyzing data such as the question sets from PeerWise makes that glaringly obvious. Since PeerWise was not developed for building exams, but rather for creating an environment for students to write and answer questions, there is a great deal of material associated with the questions that has not been considered. For example, all of the results in this thesis are from the perspective of how students perform when answering a question. This research could be linked, but currently is not, to how students perform in *authoring* questions. There are numerous ways this research could be extended. The PeerWise Community page (www.peerwise-community.org/publications) lists current academic research that either uses PeerWise data or is relevant to its academic use. A few possible areas of future work are described in the sections that follow.

5.2.1 Concept Coverage

Gauging question difficulty and question discriminating power are two essential facets of building high-quality exams. One aspect that was not covered in this work is a measure of how well an exam covers all of the topics in a curriculum. Incorporating topic coverage in exams was mentioned in the background research on concept mapping in Chapter 2. The Q-matrix approach, which automatically measures full topic coverage, would be a method worthy of further study.

The information needed for a topic-based analysis is already incorporated into the PeerWise data sets [28]. Tagging questions with one or more question topics was encouraged in the question authoring steps in PeerWise. These topics were flexible, so new ones could be added, or existing topics, including those seeded by instructors, could be associated with a new question. The approaches used in the concept mapping papers in the background section discuss the value of comparing the concept maps of students to those of their teachers.

Students who have highly overlapping concept maps with their teachers tend to perform better in the class. Because a student-authored question is recorded, I could create concept maps automatically based on the topics mentioned in the set of questions that a student authored. Perhaps the more high-quality questions that a student authors, the better he or she performs in the class. This possible correlation could be tested using algorithms mentioned in Chapter 2.1. Stochastic sampling might also provide a viable solution.

5.2.2 New Data Sets

I found a nice plateau in the cohort movement when using exams comprised of the top 15% most correlated students and the top 25% most correlated questions from the PeerWise data. Nonetheless, the data, while sufficent in terms of the number of total questions answered (62,333), still produced exams with only dozens of quality discriminating questions.

There are several experiments that the work detailed in Section 4.3.1 suggests as next steps. Testing new data sets from other courses using PeerWise might reveal different data density relationships between the students and the questions that they answered.

Also, testing data sets where the final class grades are available could be an interesting way to gauge how single exam Item Analysis compares to a semester's or year's

worth of work. The current data sets used in this thesis do not have grade information associated with the students. Data sets with student grade information are now available for future experiments on PeerWise [28].

PeerWise now provides another area of potential data mining information in the links to the student ratings of question difficulty and quality. Question quality is an interesting metric because it evokes a subtle human performance measure that attempts to answer the question: How good was this question? Goodness could mean a well-crafted question or one covering a topic the student answering the question had not previously reviewed. It might also mean something entirely different. Connections might appear when gathering the questions that students considered to be good and comparing them with questions deemed discriminating. Further, it is advantageous to have empirical results of a student's actual performance that can be compared with the student's perceived performance. This individual difficulty rating may reveal students who are aware about their failings or those who are blissfully ignorant.

The approaches described in this research were tuned to work with MCQs but their uses are far more extensible. First, many problems that do not originally appear to be MCQs can behave like them. A good example of this is one recent effort to crowd-source relief and recovery efforts after Hurricane Sandy, the most destructive hurricane of the 2012 Atlantic hurricane season [97]. Photos taken soon after the disaster by the civil air patrol were posted on a website for the general public to rate as having suffered minor, moderate, or major damage. The images were judged by thousands of people and the results helped direct emergency services in a situation otherwise lacking a comparison metric between similarly devastated communities. This approach could be viewed as a three-distractor MCQ.

Further, many microtask projects lack an internal method of rating the "goodness" of each contributor. Using Item Analysis to measure the quality of the contributors by placing them in cohorts is a possible solution to that problem. Finally, identifying when a task has been sufficiently answered is also supported by Item Analysis and the matrix-based approach for discovering stable performance cohorts. Judging sufficient completeness in microtasks allows new microtasks to be queued to the microtaskers more efficiently.

5.2.3 Omitted Questions

Students omitting questions in exams reveal a weakness in my approach, but an acceptable facet of exam building. In traditional Item Analysis, question responses are examined by cohorts and while omissions are directly part of question difficulty, they do not measure question discrimination. Question discrimination does subtly reflect how question omissions are considered in Item Analysis. In the adjacency model for building exams, every question had omissions and there were cases where there are many omitted questions.

By looking at the exams as a whole, the students who omit the most questions are relegated to the lowest-performing cohort. The students are not inherently low performing, they are simply represented by the most sparse data. Thus, instead of identifying the weakest performers, because of this unconventional method for building exams, I identified the most sparse question answerers. It would be interesting to discover how much this approach is affecting the cohort splitting method currently employed and if a manipulation of the cohort thresholds might mitigate this effect.

Another way of looking at omitted questions is to focus more on the students who are omitting the most questions and less on the questions that are being omitted. There is no measure that can add a feature to Item Analysis that corrects for omissions over incorrect answer choices. Extending the performance models of how members of the different cohorts behave to a two-step analysis might illuminate what is going on beyond question omission.

For example, consider Item Analysis as a preliminary step that allows a student to be placed in a performance cohort. The individual responses of each student on the questions he or she did answer could be analyzed to see if those responses indicate membership in a different cohort than the one currently assigned. What may result is a weighted approach that incorporates enough information from both steps to have more meaningful Item Analysis cohorts without requiring omission-free exams.

There are other resources available to deal with incomplete data in educational testing that could be incorporated into future analysis. IRT research uses algorithms that handle incomplete exam results in projections about how populations of students perform. A side-by-side comparison of the weighting- and adjacency-based approaches I use with an IRT-based method could provide valuable insight into the benefits of these algorithms.

5.2.4 Knowledge Rich Resources

Another possible improvement to my research results, especially concerning the automated answering of MCQs, is to look beyond gathering definitional data from web queries. One advantage of my approach is that it incorporates a web search shortcut ("Define: X") that restricts retrieved results to definitions. Unfortunately, college-level biology course terminology is not represented within the results of web searches. One option to circumvent this coverage issue would be to augment web data with more biology-oriented information in the term matching algorithm of my system.

Constraining the definitional data to biology-specific knowledge bases is a possible next research step. WordNet weights are a part of my system's comparison metric. A tool that uses a similar software infrastructure but added specific biological terminology references could be extremely useful. An attempt to build a "BioWordNet" based on that premise was not successful [60] but other approaches using Wikipedia-like resources appear more promising [98] [99]. The team that built IBM's Watson system noticed that "One of the primary characteristics of the Jeopardy problem is the 94.7 percent of the answers to questions are titles of some Wikipedia page" [98]. They propose a method for using Wikipedia metadata (redirects and anchor texts) to "effectively extract candidate answers from search results without a type ontology" [98].

Another recent tool appropriately scoped to the biology domain is the *Inquire Biology* project that consists of a digitized college-level biology texbook [29]. An "intelligent textbook," *Inquire Biology*, allows users to ask questions "with assistance." These assisted questions have six topic headings "define, structure, function, compare, relate, and search." The goal of this intelligent textbook is to improve the accessibility of science textbooks' content through a series of interactive features that facilitate dynamic question answering while a student is studying the textbook [29]. In a typical advanced high school or introductory college biology course, a student is expected to learn approximately five thousand concepts and several hundred thousand relationships among those concepts [100]. Linking my automated question difficulty measurement system to a resource such as this could not only improve the test preparedness of the textbook users but also help refine my research tool.

5.2.5 Minimal Exams

The smallest potential exam in terms of number of questions that contains sufficient responses for performance cohort separation consists of two questions answered by

three students. If that is the lower bound and current lengthy exams made up of excessive numbers of non discriminating questions is the upper bound, how close can exam building come to reducing the number of questions that a student must answer to prove competence? This is an important empirical issue.

A slightly more relevant question is: What is the minimal number of questions that a student must answer to adequately fit a behavior model for one of the three performance groups? This question depends on how many students take this minimal exam, but consider the efficiency of taking a 15-question test instead of a 40-question one. That difference would allow more time for students to learn instead of taking the test. Perhaps even more important would be the time saved by teachers in developing the tests.

Discovering optimal minimal exams is dependent on the number of students taking the exam, the concept coverage of the exams, and the discriminating power of the questions. Finding the balance of exam size, discriminating questions, and concept coverage attempts to maximize several variables, not unlike the constraints of building good exams to begin with. If concept topics could be accounted for and all of the questions were constructed to adequately represent the students' comprehension levels, then it would simply be a matter of finding enough students to answer those questions for significant results.

5.3 Final Summary

In summary, my thesis proposes algorithms that can answer MCQs in introductory biology with almost three times the acuracy of random guessing. My analysis is coupled with a discussion of why the bag-of-words approach, even integrated with WordNet, may have intrinsic limits in this knowledge-rich domain. Further, my thesis suggests that, with proper algorithms and analysis, crowdsourced exams can be a valuable source of exam questions that can both discriminate between student performance levels and measure difficulty levels in the questions themselves.

Appendix A

List of Questions Used as Examples

Here is a list of all of the questions used as examples in the thesis text. Some of them are also shown in Appendices G and H as they are from the PeerWise Courses 1 and 2.

Example 1

In the QA community, questions that present definitions and ask for the term being defined or ask for a word or phrase that refers to the entity or process being defined, are known as

A. Inverse Definition

B. Inverted Descriptive

C. Quiz-Style Questions

D. **All of the Above (correct answer)**

Example 2

The outward appearance (gene expression) of a particular trait in an organism is referred to as:

A. A genotype

B. **A phenotype (correct answer)**

C. An allele

D. A chromosome

Example 3

A *compound* [a] that is *synthesized by* [b] *both humans and geranium plants* [c] is known as

A. Cellulose [a]

B. **B) ATP [a+b+c]**

C. Ethyl alcohol [a+b]

D. Chlorophyll [a+b]

E. Mercury [a]

Example 4

Which hormone secretion pattern is directly affected from jet lag?

A. **Cortisol (correct answer)**

B. Insulin

C. Thyroid Hormone

D. Adrenaline

E. Calcitonin

Example 5

What is the name of the areas between osteons?

A. Canaliculi

B. Lacunae

C. Lamellae

D. **Interstitial lamellae (correct answer)**

E. Volkmann's canals

Example 6

The process that releases energy for use by the cell is known as

A. Photosynthesis

B. Aerobic metabolism

C. Anaerobic metabolism

D. **Cellular respiration (correct answer)**

E. Anabolism

Example 7

What type of **glia** *cells* engulfs and destroys *micro-organisms* and debris?

A. Astrocytes

B. **Microglia (correct answer)**

C. Ependymal *Cells*

D. Oligodendrocytes

E. Schwann*Cells*

Example 8

What hormone has a negative influence on growth?

A. Thyroid Hormone

B. Insulin

C. **Cortisol (correct answer)**

D. Testosterone/Estrogen

E. ACTH

Example 9

Which primary hormone is responsible for growth in a one-year-old child?

A. Growth Hormone

B. Cortisol

C. Insulin

D. **Thyroid Hormone (correct answer)**

E. IGF-1

Example 10

A defect in the chloride ion transporter channel is responsible for what disease?

A. Type II Albinism

B. **Cystic Fibrosis (correct answer)**

C. Wilson's Disease

D. Epilepsy

E. Neurofibromatosis

Example C.1

Molecules that are too large to pass through the pores of a cell membrane may enter the cell by a process known as

A. Hydrolysis

B. Pinocytosis

C. **Cyclosis (correct answer)**

D. Synthesis

Appendix B

Question Answering Background

This appendix is a short historical background on QA that is related to Chapter 2.

Since the question-answering track was introduced in 1999 to the Text REtrieval Conference (TREC), it has held a benchmark role in developing Information Retrieval and Question Answering techniques in the international language technology research community. Each year, different tracks are presented and dozens of universities and research institutions submit answers based on systems they build. The gold standard is manually created by the National Institute for Standards and Technology, NIST, and the subsequent evaluation of these systems has been one of the most interesting and best documented in the community. The QA systems are built with a focus on the data sets provided by NIST. These systems are then run with a set of questions that NIST provides. Questions must return actual answers, not relevant documents where the answers may be found, as in other related IR tasks. While the TREC QA competition has been running for the past decade, tasks that dealt specifically with definitional questions have been in effect since 2003. It was then that a definition track was added as part of the main task. The types of questions initially assessed were factoid questions that are "fact-based, short answer question such as 'How many calories are there in a Big Mac?'" [77]. In contrast, the new definition questions sought the most important descriptive information about 50 question topics, or targets. These include diseases ("What is TB?"), historical figures ("Who is Vlad the Impaler?"), and organizations ("What is Freddie Mac?").

Perhaps the most important QA system to date was built by the T.J. Watson research group at Yorktown Heights, New York [101]. In an overview of the Watson project, the team discusses the combination of distributed cutting-edge computing power, the multilevel search of web available resources and the text analysis necessary to generate

proper answers.

Appendix C

Discovering Definitions in Text

This appendix describes how definitions are discovered in text. It also covers using the definition patterns to filter for inverse definition questions to build the question data sets. This work is related to Section 2.2.

A task closely linked to question answering is the search for word definitions in text. Various techniques have been utilized in the realm of Information Extraction (IE). In IE, keywords and textual indicators are used to indicate definitional material. There are three main ways that definitions are discovered: The first is by looking at the patterns of words and syntax present when concepts are being described, the second is by analyzing how these patterns are distributionally spread across text, and the third is by examining networks of correlated word associations.

Google, Microsoft's Bing, and Yahoo's web query tools have all evolved beyond simple boolean searches to include shortcuts that can be used both when users type a series of words into their search windows, and via the application programming interface (API). One such shortcut limits returned results to glossaries of definitions when the user types "def:" before the words or concept whose definition is sought. While the actual algorithms used by these three search engines is unknown to the user due to trade secrets, there is related research that may direct further analysis of how definitions are presented in text and thus, how best to discover them automatically.

At TREC, the text retrieval competition and conference that has set research precedence in QA, the competing systems try to discover the most relevant facts about question topics in the definitional question track. These most relevant facts are "information nuggets" [79], also generally called nuggets in QA, and they are chunks of descriptive information that are retrieved about a topic. They usually consist of short phrases found in close proximity to the key term. The Information Extraction components of

QA systems incorporated algorithms that delineated a series of lexical patterns that are often used when an author is presenting definitional material in text. These patterns include appositives and copula constructions, propositions, relations, and structured patterns such as a rule devised by Xu:

"<TERM>,?(is|was)? Also? <RB>? called|named|known+as <NP>" [61].

When this pattern is applied to a parsed sentence, "the rule will match the question target (<TERM>), optionally followed by a comma, optionally followed by "is" or "was," optionally followed by "also," optionally followed by an adverb (<RB>), followed by "called" "named" or "known as" and followed by a noun phrase (<NP>). In the pattern, the "?" denoted optional, "+" concatenation, and "|" alternative" [61]. These are regular expression characters widely used in many programming languages. Thus, if the initial question being used for retrieval is "'What are tsunamis?', the pattern will extract the phrase 'Tsunamis, also known as tidal waves, are caused by earthquakes'" [61].

After identifying the lexical constructions that are used in text to describe a person, place, or concept, these same patterns are used in rewriting rules that incorporate the key terms in a question. Rewriting the initial question into a phrase that is likely to occur in online text and then sending those as search queries has some benefits over using just the initial query alone. There are also some downsides, because "search queries are statistically built, causing two promising lexico-syntactic clauses could be submitted in the same query, lessening the retrieval of descriptive phrases" [3]. Thus, more sophisticated uses of these definitional patterns have yielded better results.

Nuggets that contain descriptive information (and usually follow the descriptive constructions mentioned earlier) have additionally been projected onto web documents to find added support for these nuggets, or more interestingly, to find word overlaps and correlations between definitions and terms in the corpora. Using the surface patterns leads to good results in finding definitional phrases, but then looking at the deeper lexicalized dependencies, which use lexicalized trees to build definitional sentence-oriented treebanks, shows even better results. Step-wise, "correlated words were... used to form a centroid vector, so that sentences can be ranked according to the cosine distance to this vector" [3]. Next, these correlations are augmented with sets of frequently co-occurring terms from Google snippets and "descriptive sentences, taken from the Web, can be characterized by some regularities in their lexical dependency paths. These regularities are assured to identify definitions in web documents" [3]. Examples of these regularities may be seen in Figure C.1.

Where X stands for the Definiendum

Q1 = "X"

Q2 = "X is a " OR "X was a " OR "X were a " OR "X are a"

Q3 = "X is an " OR "X was an " OR "X were an " OR "X are an "

Q4 = "X is the " OR "X was the " OR "X were the " OR "X are the "

Q5 = "X has been a " OR "X has been an " OR "X has been the " OR "X have been a "

OR "X have been an " OR "X have been the "

Q6 = "X, a " OR "X, an " OR "X, the " OR "X, or "

Q7 = ("X" OR "X also " OR is " OR "X are ") AND (called OR nicknamed OR "known as")

Q8 = "X became " OR "X become " OR "X becomes "

Q9 = "X which " OR " X that " OR "X who "

Q10 = "X was born " OR "(X)"

Merging Q7 and Q10 into a new Q11:

"X also called ", "X also nicknamed", "X also known", "X is called", "X stands for",

"X is known", "X are called", "X are nicknamed", "X are known", "X was born", "X was founded",

"X is nicknamed"

Figure C.1: An example of eleven search queries based on lexico-syntactic constructions for WebQA from Figueroa [3].

It is taken as a given that definitional lexico-syntactic constructions play a prominent role in the definition seeking shortcut that is now used in search engines. The "define: X" shortcut discovers descriptive phrases in materials indexed on the Web. The shortcuts also access online reference resources such as web-based dictionaries and encyclopedias including Wikipedia (http://en.wikipedia.org), the Merriam-Webster Dictionary, the Free Dictionary, and Answers.com. These resources, coupled with search engine shortcuts, removed the need to purposely build databases of factual material for accessing definitions.

The open source online community's growth not only resulted in Wikipedia but also fostered a series of collaborations with other systems, including new statistical approaches based on this freely available and well-structured tool. Wikipedia was used by teams beginning with the TREC 2004 Question Answering track to augment existing algorithms, including Ahn et al. [102] who used the "open domain encyclopedia, both as an additional stream for answering factoid questions, and as an *importance model* to help us answer "other" questions" [102]. Importance modeling meant that the more general facts that were retrieved in their system were compared to trusted, "high-quality sources of information that model a user's ability to distinguish between important and unimportant facts" [102]. Wikipedia was also used "as it is relatively wide-coverage, its availability in a standard database format, and the fairly structured format of its entries" [102] in comparison to other online resources such as the biography pages of encyclopedias (e.g., http:biography.com) or other topic-specific knowledge bases (e.g., Internet Movie Database (http://imdb.com)). Further collaborations, such as combining Wikipedia with WordNet to create the YAGO ontology, were also used to mine definitions for QA [103].

"Def: X" is a useful cross-search engine shortcut that incorporates using trusted knowledge bases and definitional phrases to retrieve relevant definition lists. In addition, using search engines' APIs builds upon their pre-existing and extensive web page indexing and storage systems, which lessens both "the retrieval and costly processing of a wealth of documents" [3]. Using these resources as components in larger systems is a valid approach as retrieved "web snippets have proven to be promising for answering difficult queries like definition questions" [3].

The Figure C.2 shows the question types that were used to annotate the questions and answers in the answer materials. The first four types are non-factoid and the last two, in italics, are factoid types. Factoid questions have only one correct answer; this differentiates them from list questions, which seek multiple answers. This chart is an

elaboration on the types presented by Quateroni [4].

Question Type	Explanation	Example
LIST	A list of items	"What were Columbus' three ships?"
DEFINITION	A definition or description	"What is platinum?"
HOW	An explanation	"How did Socrates die?"
WHY	A generic cause	"Why does the moon turn orange?"
OBJECT	A generic entity	"What is Grenada's main commodity export?"
SYMBOL	A visual representation or theorem	"Which formula represents an organic compound?"

Figure C.2: Question type, explanation, and an example using the Quateroni types [4].

After grouping the questions and answers into their types, the largest group of questions consisted of the object and definition types. The object and definition questions were then grouped based on the lexical patterns found in the questions. Many of the object and definition questions were in the "which" format, but the largest group gave a definition and sought a slot-filler that named the term described. That group, although similar to the TREC "definition" and "other" question, was the inverse. Hence, questions that present definitions and seek the term defined as the target are called "inverse definition questions."

In an effort to further isolate the questions that are the most lexically similar of the inverse definition questions, a subset of 69 object-definition questions were filtered into two groups: Wh-questions and inverse definition questions. After the filtering, there are 30 Wh-questions and 39 inverse definition questions. Inverse definition questions are approximately 57% of the object and definition questions and approximately 31% of the 125 questions in the initial test set.

The inverse definition questions were then grouped by their structure. A linguist examined the phrases that occurred immediately before the end of the questions. In Example C.1, that phrase is "known as." The linguist grouped the phrases by word classes: verbs and nouns. Two of the verb variants, "(also) known as" and "is (also) called" were used as the patterns for creating the test data set. They were the most frequent and self-similar type of the structural variants and are listed in Figure C.3.

Verbs	Nouns
(also) known as	function is
is (also) called	procedure is
	factor is
is found in	concept that
involved in	are organisms that
connected by	example of
function as	principle of
is termed	by the process of
is referred to as	in the form of
made up of	activities controlled by

Figure C.3: Variants in the inverse definition question set as verbs and nouns.

There are many phrases used in the question set that are close in meaning to "also known as" and "also-called" and they were included. An example question:

Example C.1

Molecules that are too large to pass through the pores of a cell membrane may enter the cell by a process known as

A. Hydrolysis
B. Pinocytosis
C. **Cyclosis (correct answer)**
D. Synthesis

Example C.1 is included as a question in the inverse definition development set. The earlier example, Example C.1, has the close variant structure "is referred to as" is also grouped as an inverse definition question.

After analyzing the Biology questions I performed the same series of analysis, typing, and lexical filtering on the psychology question set and chose 20 questions from this domain to use for the subsequent experiments. Two lexical patterns were used, "also-known-as" and "also-called," to separate these inverse definition questions into a grouping of 59 questions (39 from the biology domain and 20 from psychology), that were used for the early experiments. These questions are called the inverse definition development set and were used for early experiments on how best to correctly answer and measure difficulty of IDQs. The first step for judging the difficulty or discriminating power of questions is to build exams on which Item Analysis can be performed. The process of turning sets of questions into new exams is described in Section 3.3.

Appendix D

Adjacency Matrices for Exam Building

This appendix gathers together support documentation on different aspects of exam building. In the first section is a description of how exams are built from sets of questions answered by the same students using an adjacency matrix approach. Then, an algorithm and an example solution to the NP-hard problem of exam building is presented. Next, there is a discussion on the implementation of exam building and finally a reflection on the role that omitted questions play in this approach.

Is there an Exam, or Sets of Exams in the Data?

Since my Item Analysis algorithm depends on splitting the group of students who took the test into three subgroups, I need the scores and student set size to be sufficiently large. Furthermore, the sample data has many omissions, as students choose which questions they want to try answering.

This 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) [94] as shown in Figure D.2. I am seeking a good set that is similar to an exam. By using a heat map in which correlated data appears as darkened images to show the group of students who have answered the same questions (Figure D.3), I am presented with a realistic exam in which there are a few holes for omitted questions. A heat map is a method of visualizing the density of data through colors (or in black-and-white versions, darkness) that mirror those found in flames. Darker colors equate to denser data, and in the case of a sparse matrix, can indicate regions of denser information. In exams, students often fail to answer all of the questions and these are referred to as omitted questions. These omitted questions would be missing edges in Figure D.2. The heat map presented in Figure D.1 shows the data sorted to reveal the most dense group of students who have answered the same

question. It also allows further analysis of this dense region to discover a maximal graph, or exams with no questions omitted.

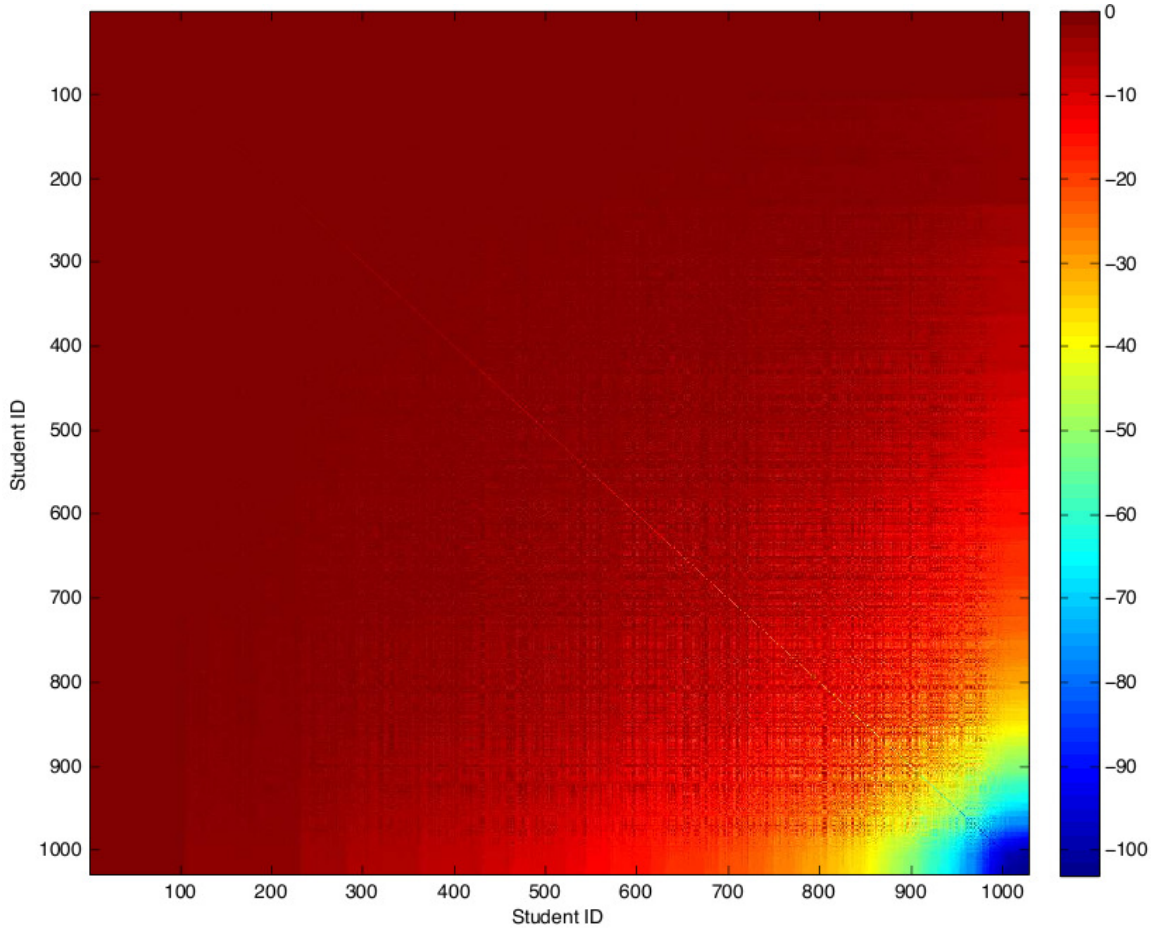

Figure D.1: Heat map of the covariance matrix for data Set1, based on the number of students who answered the same questions. The x-axis orders the students by who answered the most questions multiplied by those students' transpose. The y-axis is those students' transpose. Again, dark red represents uncorrelated pairs, whereas blue represents correlated pairs.

Finding a biclique in a larger semi-definite correlation matrix is an NP-hard problem [92]. Discovering the single maximal clique is the ideal scenario but in this situation, I only need to find a sufficiently large clique. Seeking the set of students who have answered the same questions would mean comparing each student's questions to the questions answered by all other students, pairwise and iteratively. That is an NP-hard problem and discussed further below.

Again, the steps for building and sorting the covariance matrices were introduced in Chapter 3, and are as follows:

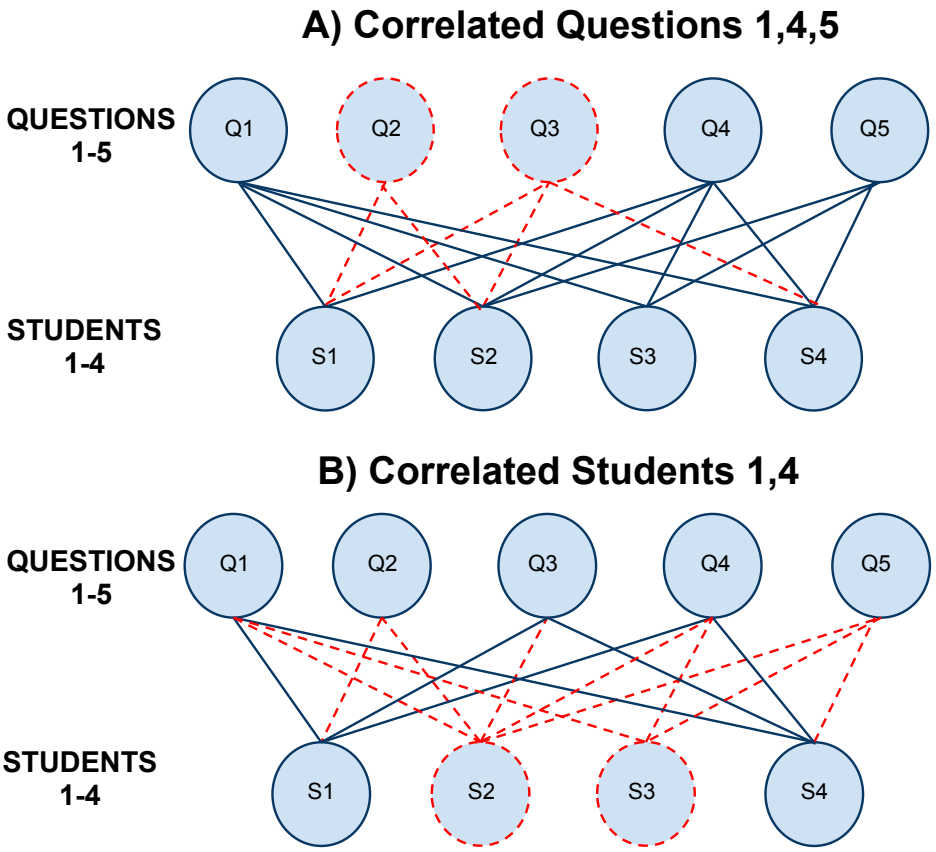

Figure D.2: Correlated questions and students as connected cliques in a bipartite subgraph. The edges represent each unique question-student pair that is recorded every time a student answers a question. In the top graph, the solid edges belong to the most correlated questions and, in the bottom one, the solid edges belong to the most correlated students.

1. Collect the data in triples of student ID, question ID, and answer choice.

2. The students are ordered by the number of questions they answered.

3. Build the incidence matrix M, with students corresponding to rows and the questions to columns. If a student answered a question, a 1 is placed in the appropriate column, if they did not, a 0 is placed in the space. The incidence matrix in Figure D.6 (1) is the bipartite graph shown in Figure D.2.

4. Compute $S = M \times M^T$. A heat map of S can be seen in Figure D.3.

5. Compute $Q = M^T \times M$.

6. Find the most correlated students by computing the vector s by summing over the rows of S, thus, $s = \sum_j S_{ij}$. Then sort the rows and columns of S based on the ordering of s because S is symmetric. This effect can be seen in Figure D.1.

7. As above, find the most correlated students by computing the vector $q = \sum_i Q_{ij}$. Then sort the rows and columns of the matrix Q based on the ordering of q.

This enumerated process produces results that are shown in heat map Figure D.2. To reiterate, the goal is to select the most highly correlated students and questions in order to build new exams and perform Item Analysis on them.

Next an example of this algorithm is presented. In Figure D.5 and Figure D.6, each question was given an identifier from 1 to 5. Each student, of which there were four, was given an identifier from 1 to 4. An incidence matrix M of size 5×4 was generated in which each row corresponds to a student and each column to a question. If a student answered a question, a 1 was entered into the incidence matrix at the appropriate row and column. All of the other spaces contained 0s.

M^T is the matrix transpose. Transposition is the interchange of the matrix where the value in row i is moved to column i.

Given that M looks like Figure D.5 (1), the transpose of M, M^T, looks like Figure D.5 (2). Multiplying the matrix M^T with M produces a covariance matrix C of 5×5, the sum of which reveals the most correlated questions. Each cell of the covariance matrix contains the "correlation index" Cij that is a metric of how well correlated sentence i is with sentence j, which is shown in Figure D.6 (1). The sum of the rows in $M \times M^T$ is of size 4×4 and presents the most correlated students, as seen in Figure D.6 (2).

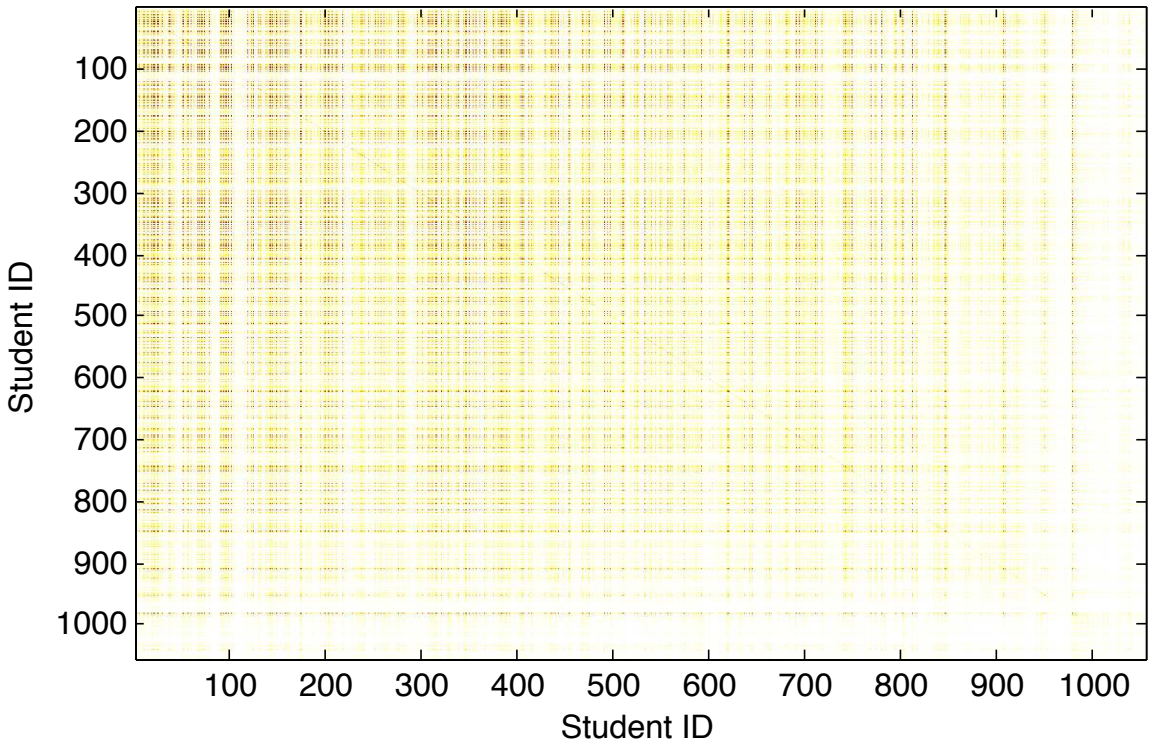

Figure D.3: Heat map of the correlated students before they are sorted to reflect the most correlated sets. Here, the white represents uncorrelated pairs and the black shows correlated pairs. The pattern reflects the sparse information areas found in the uncorrelated data.

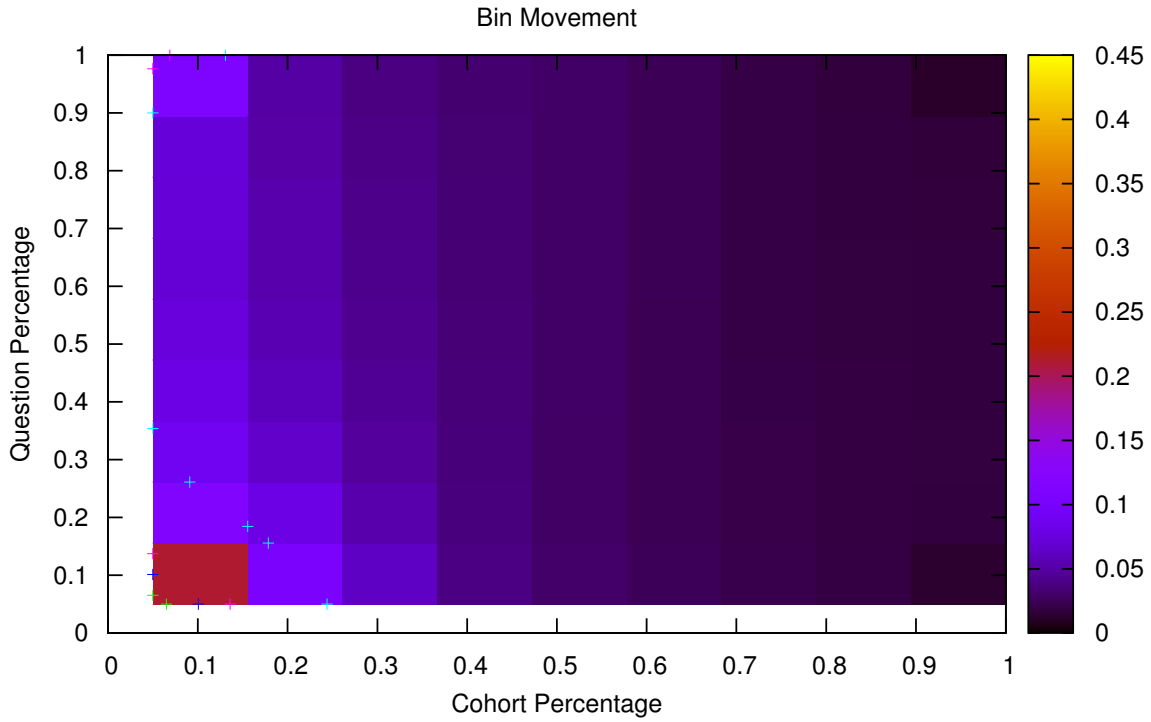

Figure D.4: A heatmap presenting the movement of students to different performance bins based on the percentage of most correlated students and questions. Both correlated students and questions are shown from 0.05% to 100%. This data was gathered before omissions were taken into consideration, so any number of omissions are permitted.

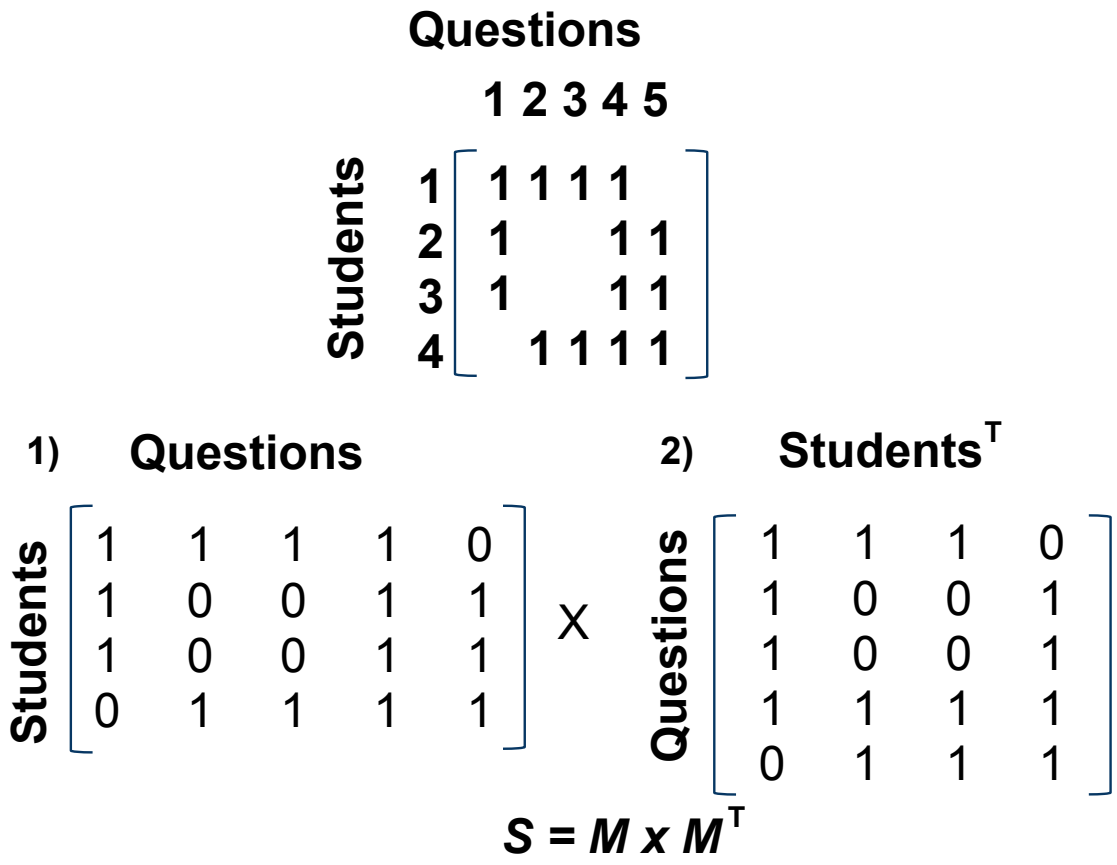

Figure D.5: Setting up the correlation matrix where ones represent a student answering a question and zeros are question omissions. *S* is the most correlated students and *M* is the matrix of the students and questions that is used in the matrix multiplication.

1) **C (most correlated questions) =** **$M^T M$**

$$\begin{bmatrix} 3 & 1 & 1 & 3 & 2 \\ 1 & 2 & 2 & 2 & 1 \\ 1 & 2 & 2 & 2 & 1 \\ 3 & 2 & 2 & 4 & 3 \\ 2 & 1 & 1 & 3 & 3 \end{bmatrix}$$

$$\boldsymbol{10 \quad 8 \quad 8 \quad 14 \quad 10}$$

or Question Numbers 1, 4, 5 (also shown in Figure D.2)

2) **C^T (most correlated students) =** **$M M^T$**

$$\begin{bmatrix} 4 & 2 & 2 & 3 \\ 2 & 3 & 3 & 2 \\ 2 & 3 & 3 & 2 \\ 3 & 2 & 2 & 4 \end{bmatrix}$$

$$\boldsymbol{11 \quad 10 \quad 10 \quad 11}$$

or Student Numbers 1, 4 (also shown in Figure D.2)

Figure D.6: In the correlation matrix, using transpose and sum to reveal the most associated questions and students. The most correlated questions are shown in Part 1 and correspond to Part A of Figure D.2. The most correlated students are shown in Part 2 and correspond to Part B of Figure D.2. Correlation matrices represent the same information as connected cliques in bipartite subgraphs, which are shown in Figure D.2.

This graph-based algorithm at its essence prioritizes the set of students and questions that should be searched first to create the optimal desired exam. A user may seek multiple sets of exams with a varying balance between the number of students and the number of questions. For example, one may seek an exam with a large number of students and a small number of questions in an effort to give a test that quickly differentiates students into performance cohorts via Item Analysis. As noted in Section 3.3, an exam must contain enough students to permit the students to be split into three performance groups: low, middle, and high, based on their exam score.

This sorting process provides a sound heuristic for selecting highly correlated students and questions. I then select 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 based on the adjacency methodology. This presents a realistic exam where there are a few holes, i.e., omitted questions, or missing edges in Figure D.2. Once I have built the exam, I move onto analysis of the individual question difficulty and discrimination power.

Complexity Issues

Next is a discussion of the completeness issues related to this research. The individual student question answering relationship is represented with a graph: an "exam" in which every student answers every question is represented by a complete bipartite graph (or biclique). I seek a good set that is similar to an exam.

Non-deterministic polynomial-time hard, or NP-hard, means that as a set of data gets large, the time needed to solve a problem is prohibitive due to the large number of computations required [92]. In essence, as the numerical value of the two variables increases, the problem becomes uncomputable. Since the students and questions can be represented as a graph whose edges are traversed to discover if any questions share membership in a clique comprised of a set of students, the pair-wise comparisons needed to answer this problem get extremely large as the numbers of students and questions increase.

Consider the problem of finding a sufficiently large set of the same students who have answered the same questions in an exam as that of comparing edges as in Figure D.2. 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 he or she has also answered the first question.

In Figure D.2, S4 or student 4 is the only student to have answered the same ques-

tions as S1. This iterative checking of group membership is exacerbated by the fact 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. With two data sets used in this work containing in Set1:

- 148 questions
- 1055 students
- 31,019 shared edges between the questions and the students

and in Set2:

- 134 questions
- 887 students
- 31,314 shared edges between them.

This exam building is an attempt to solve the "Zarankiewicz problem" [104] to determine "the minimal number of edges in a bipartite graph which guarantees the existence of a complete balanced bipartite subgraph Kq,q." "Finding a largest balanced complete bipartite subgraph is an important optimization problem, which is known to be NP-hard, and even hard to approximate" [104]. There are instances where the solution is tractable [105] and [104]. In certain large data situations however, the solution may not be tractable. Since I seek a solution that is useful for data sets of varying dimensions that may not have "constant density," the adjacency matrix approach gets a close approximation.

Building Exams by Extending the Adjacency Matrices

There are two approaches for building the exams: either maximize for students or maximize for questions. In other words, either search for students who share questions or questions that share students. The end goal is the same: Find the largest exam to then support experiments downstream and related analyses, such that those experiments and analyses are based on a sufficiently large data set to be statistically significant. Two competing variables are attempting to be maximize when creating a large exam: questions and students.

A priori, there is not an exact number of questions or students that is ideal. There could be multiple, viable, and discrete exams built from a given data set. Since this research deals with test item difficulty, the minimum exam must contain at least three students that split into the three performance cohorts, for the subsequent Item Analysis to yield meaningful results. There are several scenarios where three students would not split into different cohorts, such as a situation where all three students are high achieving and answer all questions correctly, or a situation where the questions lack discriminating power. To eliminate obviously non discriminating questions, questions are sought where the whole set of students answering either all got the question correct or all got it incorrect. There were no instances of this behavior in either data set.

Again, the ideal exam size is the largest set of questions answered by the largest set of the same students. The search for sufficiently large maximal question-student sets via adjacency matrices uses the addition of columns in covariance matrices to avoid the brute-force method. This brute-force method entails iteratively searching through the questions and tests whether two students have answered one question in common, and then individually testing the following question for commonality, one-by-one. By summing the columns of the covariance matrices presented in Section 3.4.1, students are clustered who share a number of questions in common.

As discussed in Section 3.4.1, S_{ij} is the number of questions that student i has in common with student j. In the case where i is the same as j, this is the number of questions a student answered. Q_{ij} is the number of times that a student answered both question i and j. In the case where i is equal to j, this is the number of times a question was answered. These are symmetric covariance matrices but the items in the matrix are not based on a regular scale like age or weight. These are sets of questions that have been answered by groups of students.

To reiterate, for the purposes of performing Item Analysis the minimum number of students and questions would be three students and at least two questions. For this minimum-sized exam, the students' performance would also have to fall into the three achievement groups. One student would need to get both questions correct, one would need to get both incorrect and the third would need to get one correct and one incorrect. This minimally sized new exam would force a layout similar to the exam on the left in Figure D.7. Another example of a minimum-sized exam with three questions is shown on the right.

In this example, the two questions, Q1 and Q2, are answered by three students: S1, S2, and S3. Zeros represent incorrect answers and ones correct answers. This is one

	Q1	Q2
S1	0	0
S2	0	1
S3	1	1

3 Student × 2 Question Exam

	Q1	Q2	Q3
S1	1	0	0
S2	0	1	1
S3	1	1	1

3 Student × 3 Question Exam

Figure D.7: Two examples of minimum-sized exams sufficient for Item Analysis. The zeros represent incorrect answers and the ones correct answers. The students are S1, S2, and S3 and the questions are Q1, Q2, and in the case on the right, Q3.

acceptable minimum exam but there are several possible answer correctness variations on this new exam that would not produce three distinct cohorts for Item Analysis. This motivates new exams comprised of many students who have answered the same many questions. While there is a minimum exam size for Item Analysis, there is no maximum size. National and international standardized exams are taken by hundreds of thousands of students a year.

If every student got at least one question correct, a 3×3 matrix might work, but the differences between the student performance must still be measurable as in Figure D.7. Thus, the lower bound on new exam size is that Q must be greater than 1 and S must be greater than or equal to 3.

Reflections on Omitting Questions

The sufficiently large clique model described for building exams seeks to maximize for questions answered by the same students. In fact, in actual exams, students often skip questions. This occurs because students either lack sufficient time to answer all of the questions in an exam, or because in some grading formats students are penalized for wrong answers and are consequently encouraged to leave questions unanswered that they are guessing with a high level of uncertainty. Finally, there are also simple mistakes made during exams, which also result in questions being left unanswered. Gronlund and Davis take this student behavior into consideration in their descriptions of how to build achievement tests that measure performance of students [36] [76]. Con-

sequently, they include a row for omissions (row 7, just above "total") in calculating Item Analysis, see Figure 2.6.

As a result, the new exams may be built to incorporate various degrees of question omission to replicate the type of behaviors that actual students exhibit while exam taking. This is an elective choice, since the original exam creation goal was to identify the most dense matrix of students and questions. Allowing exams to include some question omissions mirrors actual exam behavior and increases the pool of acceptable students in the newly generated exam. Following the lead of Gronlund and Davis, I initially sought an exam of 100% participation by the students. Although it might seem that 100% participation would be ideal, our empirical evidence indicates that omission is not the best predictor of question difficulty or discrimination. The data shows that the omission rate is not as illuminating as I earlier had thought. Thus, I sought an approximation of an exam with 100% participation. This loosened approach to exam building, increased the newly created exam size substantially.

Allowing the newly built exams to contain students who have answered most, but not all the questions, as students often do in natural test-taking, would increase potential exam sizes. Discovering what percentage of questions that may be omitted in an exam without negatively affecting the statistical significance of the cohort groupings increases the number of students whose questions can be included in an exam. This will greatly increase the amount of viable exam data provided by this adjacency matrix approach.

Appendix E

MySQL Data Characteristic Queries

Here are the MySQL queries associated with the data characteristics described in Section 3.3.2. The list presents the observation in italics followed by the relevant MySQL query:

- *Data Set1 comprised 1055 students and 148 questions split into two subgroups.*
- *Data Set2 comprised 887 students and 132 questions split into two subgroups.*
- mysql> select distinct(user) from answers where course='1'; [number of students in course 1 or 2]
- mysql> select distinct(question_id) from answers where course = '1' [number of questions in as course 1 or 2]
- *The least number of questions answered by any of the students was 1. 101 students in Set1 only answered 1 question; 152 students in Set2 only answered 1 question.*
- mysql> select s.question_id, count(*) as c from answers as s, author_answers a where s.question_id = a.question_id group by s.question_id order by c desc;
- *112 was the most questions answered by a student in Set1; 11 students answered 112 questions.*
- mysql> select s.question_id, count(*) as c from answers as s, author_answers a where s.question_id = a.question_id group by s.question_id order by c;
- *The average number of questions answered by students in Set1 was 26.6 and in Set2 was 35.8.*

- mysql> select distinct(user) from answers where course='1'; [number of students in course divided by the number of answers:]
- mysql> select count(*) from answers;
- *None of the students answered any question more than once.*
- *None of the questions were so easy that all of the students answered them correctly, nor so hard that none of the students got them correct.*
- *Set1 contained 31,019 answers and Set2 had 31,314 answers.*
- mysql> select count(*) from answers;
- *The most times a single question was answered was 439; the least was 89.*
- mysql> select distinct(question_id), count(*) as c from answers where course = '1' group by question_id order by c desc limit 1;
- mysql> select distinct(question_id), count(*) as c from answers where course = '1' group by question_id order by c limit 1;
- *In Set2 331 was the maximum number of times a question was answered; 132 was the minimum.*
- mysql> select distinct(question_id), count(*) as c from answers where course = '2' group by question_id order by c desc limit 1;
- mysql> select distinct(question_id), count(*) as c from answers where course = '2' group by question_id order by c limit 1;
- *There are 62,333 distinct answers or questions that were answered in total.*
- mysql> select count(*) from answers where course = '1'; [31019]
- mysql> select count(*) from answers where course = '2'; [31314]
- *20,532 of the total answers by students were incorrect, or 32.7%.*
- mysql> select count(*) from answers [minus the line below, then, total number incorrect divided by total number of answers]
- mysql> select count(*) from answers s, author_answers a where s.question_id = a.question_id and s.choice = a.answer;

- mysql> select count(*) from answers s, author_answers a where s.question_id = a.question_id and s.choice != a.answer;
- *There were 14,094 question ratings, and each of the 280 total questions from both sets were rated at least once.*
- mysql> select distinct(question_id) from ratings;

Appendix F

Automatic Answering Pipeline

This appendix covers the details of the automatic question analysis pipeline presented in overview in Section 3.2.

After the question sets are broken up into uniform groups, the groups are automatically marked up in XML with a Python script. This script also removes any italics or unusual characters that may cause problems later in the pipeline.

Here is an example question before processing:

Example 6

The process that releases energy for use by the cell is known as

A. Photosynthesis
B. Aerobic metabolism
C. Anaerobic metabolism
D. **Cellular respiration (correct answer)**
E. Anabolism

The Python script transforms Example 6 into the format shown in Figure F.1 for the next stage of processing. The answer options are shown in all capital letters in Figure F.1 so that they are more visible to read in this document. The structure of the XML creates a primary structure or "backbone" for future processing. When the answer options are sent to the search engines in the following pipeline step, the results are piped back into the coordinating <ResultSet> tags, as noted in Figure F.2 and later shown in Figure F.3.

System Implementation

The entire system implementation is shown in Figure 3.4. Each major section of the data flow pipeline is shown left to right starting with the input described in Section

```
<test description="test">
<section description="5">
<question Group>
<question ID="1" correct="D">
<text>The process that releases energy for use by the cell
is known as
</text>
<answer id="A"><text>PHOTOSYNTHESIS</text>
<define id="A">
<ResultSet>
</ResultSet>
</define>
</answer>
<answer id="B"><text>AEROBIC METABOLISM</text>
<define id="B">
<ResultSet>
</ResultSet>
</define>
</answer>
<answer id="C"><text>ANAEROBIC METABOLISM</text>
<define id="C">
<ResultSet>
</ResultSet>
</define>
</answer>
<answer id="D"><text>CELLULAR RESPIRATION</text>
<define id="D">
<ResultSet>
</ResultSet>
</define>
</answer>
<answer id="E"><text>ANABOLISM</text>
<define id="E">
<ResultSet>
.
.
.
</ResultSet>
</define>
</answer>
```

Figure F.1: Question format after preprocessing the contents of Example 6.

```
<answer id="D"><text>cellular respiration</text>
          <define id="D">
<ResultSet>
THE 50 RETRIEVED RESULTS FOR "DEFINE: CELLULAR RESPIRATION"
WILL BE PLACED HERE
</ResultSet>
</define>
</answer>
```

Figure F.2: Question format indicating the retrieved results storage location.

3.2.1. The 50 question documents contain 20 questions each and are marked up in XML. These question documents are used as input to the XProc-based query system, which identifies the IDQ and the related answer options, sends the answer options to a search engine with the prefix "Define:," and post-processes the stored results into a version that is input to the ROUGE comparison system [5].

I used Calabash [89], an open source implementation of XProc written in Java. Building an XML pipeline supports data traceability and in the entire pipeline the original retrieved information from the queries is retained. The specific format used by ROUGE is 251 individual documents per question, one for the initial question and 250 for the retrieved result titles and snippets. The output of the ROUGE system is a results document that is passed to the post-processing component of the pipeline. In this component, the results are merged, filtered, and compared to the actual correct answer. Then, the original XML question document is augmented with this information. Finally, more extensive analysis is provided by using the Marklogic XML database [88].

The XProc Query System

Once the original questions have been turned into XML documents, those documents become the input to the XProc [90] pipeline which in turn carries out a series of queries and transformations based on the XML tags associated with the data structure of those documents. The input of this system is shown in Figure F.2. Then, a second XProc pipeline processes the output of these queries into a format accepted by ROUGE (described in Section 3.2.2).

XProc is an XML-based pipeline language that describes transformations to be performed on XML documents. XProc is particularly suited to this definition-seeking task, because beyond the actual queries to search engines, the data processing can be

viewed as a series of manipulations of tagged information within XML documents. That is, the original input XML question set is enriched with additional data at each step in the two distinct XProc pipelines described in this section.

Figure F.3 shows the first XProc pipeline; a sub-pipeline MakeRequest.xsl is listed in Figure F.4. This initial pipeline accomplishes the following tasks:

- Each XML question document is input at the top and identifies ("viewport match") the answer options in step (1) and sends those options as the input to the main sub-pipeline.
- An XSLT stylesheet is applied to the main pipeline that make "c:request" elements in the document being processed.
- The second viewport matches those "c:request" elements to the answer options and passes them to its sub-pipeline.
- The "http-request" step sends each answer option to the Yahoo! search engine, along with the prefix "Define:"
- The XML output from Yahoo! becomes the input for the second XSLT "TagReduction.xsl" stylesheet that performs text manipulation on the retained titles and snippets for the top 50 results associated with each answer option.

The XML output of this pipeline hides unnecessary information, but never deletes any part of the query result. This methodology allows the results to be meaningfully viewed by the human eye, but refrains from removing information that might be useful at a later date for additional processing or analysis. An example of this output is seen in Figure 5.6.

Figure F.4, presents the XProc code of MakeRequest.xsl used in the iterative looping query process shown in summary in Figure 3.4. The XProc stylesheet matches the XML tag "answer/text" and sends that information, preceded with "Define:," to the Yahoo! search engine via its API. Then, the results are added to the "ResultSet" tag in the original XML document, which is renamed "results" in the process. In Figure F.4, the code documentation is shown in all capital letters.

Yahoo! was the first search engine accessed because of its standard API. The APIs for other search engines and web resources such as Wikipedia and Wolfram Alpha are similar and require minor variations of this code, but they depend on the same overriding methodology to successfully retrieve results. Other APIs, such as the one

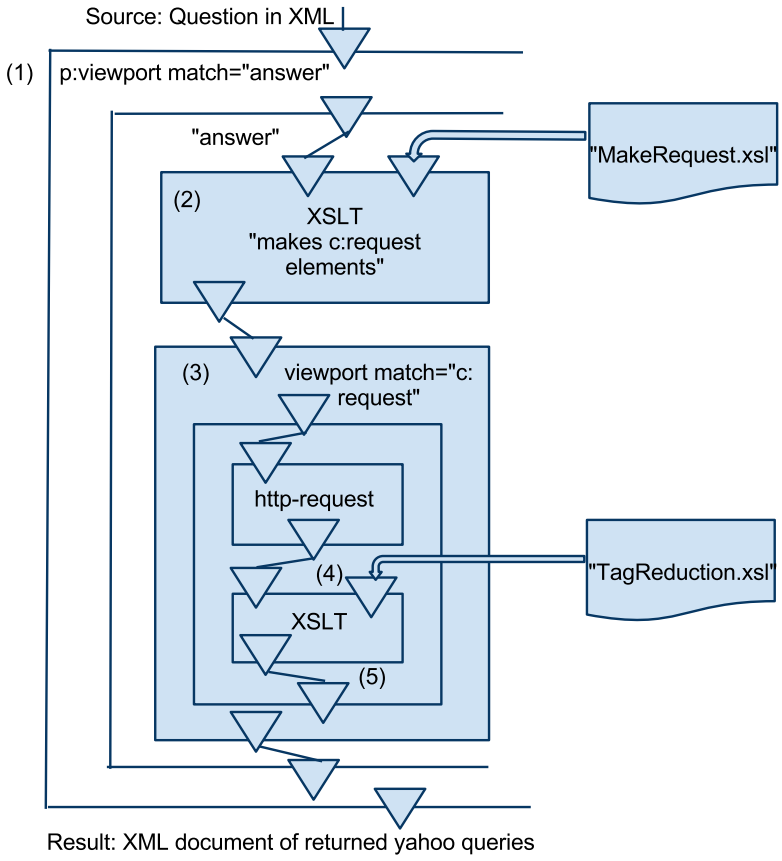

Figure F.3: The data flow of the XProc query system.

for accessing Google results, use fixed-point looping and output JSON documents, which would require additional processing on both the input and output of the queries.

Figure 3.4 presents a high-level view of how the XProc Query pipeline fits into the data flow of the entire system, Figure F.3 presents the pipeline structure of the XProc Query code, Figure F.4 shows the actual code that queries Yahoo!, and Figure F.5 presents the retrieved results. Figure F.4 shows the XProc code of MakeRequest.xsl that is noted in step (2) in Figure F.3. The results shown in Figure F.5 are processed to remove the time stamp, the actual web page URL, and other information. Again, this data has been retained in the question backbone document in XML, but only the text within the <title> and <summary> tags (within the <result> tag) are the strings that will be used for further comparison in Section 3.2.2.

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- Document   : make-request.xsl-->
<xsl:transform version="2.0" xmlns:xsl="http://www.w3.org/1999
                                                /XSL/Transform">
<xsl:template match="answer/text">
 <xsl:copy-of select="."/>
    <define>
    <c:request xmlns:c="http://www.w3.org/ns/xproc-step"
        method="GET"href="http://api.search.Yahoo!.com/
        WebSearchService/V1/webSearch?appid=KtU8kZjV34GLEMp
        00tt2mGgX7r5wTBRxZ6uW4djRmEasSjVPRLPjUv3geWogNQ--
        &query=define+{encode-for-uri(.)}&
        type=all&results=50&output=xml"detailed="false"/>
    </define>
    </xsl:template>
<!--THIS MATCHES ANSWER DEFINE AND COPIES THIS INTO THE
    RESULTSET NODE, REMOVING RESULTSET-->
<xsl:template match='answer/define'/>
<!--THIS MATCHES THE ANSWER DEFINE BRANCH AND REMOVES IT
    IF THERE IS NOT ANYTHING IN THE ANSWER/DEFINE BRANCH-->
<xsl:template match="node() | @*">
<xsl:copy>
<xsl:apply-templates select="node()|@*"/>
</xsl:copy>
</xsl:template>
</xsl:transform>
```

Figure F.4: XProc pipeline code example from makeRequest.xsl that shows the query construction.

In Figure F.5, the <answer id> tag identifies the question's answer option (A-E). This tag is followed by a <text> tag that wraps the literal answer option. Next, all of the retrieved results are listed within the <define/results> tags, which is a renaming of the <Define id> attribute of the <ResultSet> tag introduced in Figure F.1. Then, this answer option is one of five in the entire question and the document is closed with a </question> tag. Finally, the retrieved results are normalized and sent on the next step in the pipeline, the ROUGE summary comparison system.

ROUGE

ROUGE is a Perl-based text comparison system that examines the resemblances between one text string and a set of other possibly matching text strings. ROUGE judges the similarity of these strings on their statistical word overlap. As a consequence it is an appropriate tool for measuring the similarity of inverse definitions to definitions retrieved from the web. The development of ROUGE was initially motivated by the very expensive cost of human judgments in distinguishing good summaries from less viable ones. Human evaluators are used in many tasks in natural language processing and as a result ROUGE, as a replacement tool, has found utility beyond summarization.

ROUGE has also been used in comparing the output of machine translation programs and the task-based evaluation showed results similar to those of humans [91]. In this case, ROUGE aided measuring the similarity between a "candidate translation and a set of reference translations" with matching the Longest Common Subsequence (LCS) and utilizing skip-bigrams [91]. LCS identifies the "longest co-occurring in-sequence n-grams"; skip-bigrams are "any pair of words in their sentence order" [91]. Using ROUGE for comparing the output of queries from the Web also benefits from automatically identifying "sentence-level structural similarity" as the text being compared to the original question is a concatenation of the title of the web page and the snippet (or summary) of the content on that page that contains the most relevant information to the initial query [91].

ROUGE uses n-gram recall between the given inverse definition and a set of reference strings (the retrieved results). Figure F.6 shows how ROUGE-N is computed, where n stands for the length of the n -gram $(gram_{n)}$ and $Count_{match}$ ($gram_n$) is the maximum number of n -grams co-occurring in the original question and the set of retrieved results [5]. ROUGE is a recall-based "measure because the denominator of the equation is the total sum of the number of n-grams occurring" in the retrieved results [5].

Once the "Define: X" definition queries have been run, the retrieved titles and

```
<answer id="D"><text>cellular respiration</text><define>
<results>
<result>
<title>Cellular respiration - Wikipedia, the free encyclopedia
</title><summary>Cellular respiration, also known as 'oxidative
metabolism', is one of the key ways a cell ... glucose molecule
during cellular respiration (2 from glycolysis, 2 ...
</summary>
</result>
<result>
<title>Cellular respiration - Definition</title>
<summary>Cellular respiration is, in its broadest definition,
the process in ... In cellular respiration, this oxidation
process is broken down into two basic metabolic ...
</summary>
</result>
<result>
<title>cellular respiration: Definition from Answers.com
</title>
<summary>cellular respiration n. The series of metabolic
processes by which living cells produce energy through the
oxidation of organic
</summary>
</result>
<result>
<title>cellular respiration - Medical Definition</title>
<summary>Definition of cellular respiration from The American
Heritage Medical Dictionary.
</summary>
</result>
<result>
<title>Cellular respiration - definition from Biology-Online.org
</title>
<summary>Definition and other additional information on
Cellular respiration from Biology-Online.org dictionary.
</summary>
</result>
<result>
<title>CELL RESPIRATION.doc</title>
<summary>What organelle in the cell carries out
cellular respiration? ... What is the definition of cellular
respiration? process that releases energy by breaking down
food molecules ...
</summary>
</result>
</results>
</define>
</answer>
```


Figure F.5: The first five results and result 12 from "Define: Cellular Respiration."

$$\text{ROUGE} - \text{N} = \frac{\sum_{S \subset \{\text{ReferenceSummaries}\}} \sum_{gram_n \subset S} Count_{match}(gram_n)}{\sum_{S \subset \{\text{ReferenceSummaries}\}} \sum_{gram_n \subset S} Count(gram_n)}$$

Figure F.6: How ROUGE-N is computed from [5].

snippets are sent to a new folder because of their large size. These titles and snippets may include the definition being sought because of the "define" shortcut. This shortcut uses the search engine's own metrics to associate the queries specifically to strings of text that make up definitions. Thus, the retrieved title and snippet are used to assess whether the retrieved definitions are approximations of the inverse definition.

This process may produce a great deal of data files that may pose additional management challenges. The folder for each complete pipeline run contains 1000 folders, one for each question. Every question folder contains 251 documents formatted with one sentence per line or ".spl," which is the hardcoded document format that ROUGE uses. This includes the model path, or the original question, and 250 peer paths (retrieved string made up of the text of the title and snippet) in the form seen in Figure F.7.

```
FILE STRUCTURE  ROUGE NUMBERING
/X
model_pathX.spl
peer_pathA_1-50.spl 1-50
peer_pathB_1-50.spl 51-100
peer_pathC_1-50.spl 101-150
peer_pathD_1-50.spl 151-200
peer_pathE_1-50.spl  201-250
```

Figure F.7: The structure of the number files, where "X" is the question number and the number that it corresponds to in the ROUGE results.

Next, the contents of each question number folder are processed by a script that runs ROUGE and then changes the contents of the ROUGE configuration file before running the next question. Because ROUGE has hard-coded input and output files this script is a work-around for iterating through the numbers, changing the input files, running ROUGE, getting output, and then renaming the output to avoid being written over by the next use. Figure F.8 shows the textual content of the structure shown in Figure F.7 with the first sentence (1) being model_path82.spl and the following five sentences

(D_1-5) being the first five retrieved results for "Define: Cellular Respiration." The last sentence (D_12) is the twelfth retrieved result and also the top scoring result for the entire question. "Cellular respiration" is indeed the correct answer to the question presented earlier in Example 6.

> The process that releases energy for use by the cell is known as
>
> D_1) Cellular respiration - Wikipedia, the free encyclopedia Cellular respiration, also known as 'oxidative metabolism', is one of the key ways a cell ... glucose molecule during cellular respiration (2 from glycolysis, 2 ...
>
> D_2) Cellular respiration - Definition Cellular respiration is, in its broadest definition, the process in ... In cellular respiration, this oxidation process is broken down into two basic metabolic ...
>
> D_3) cellular respiration: Definition from Answers.com cellular respiration n. The series of metabolic processes by which living cells produce energy through the oxidation of organic
>
> D_4) cellular respiration - Medical Definition Definition of cellular respiration from The American Heritage Medical Dictionary.
>
> D_5) Cellular respiration - definition from Biology-Online.org Definition and other additional information on Cellular respiration from Biology-Online.org dictionary. ...
>
> D_12) CELL RESPIRATION.doc What organelle in the cell carries out cellular respiration? ... What is the definition of cellular respiration? process that releases energy by breaking down food molecules ...

Figure F.8: The first five and twelfth retrieved results for "Define: Cellular Respiration." The format of this figure is how the returned results appear from an answer option in Example 6.

ROUGE is a flexible tool that allows different metrics for comparing text strings. My baseline ROUGE implementation, which sought the unigram match of the literal word overlap between the original IDQ text and the text of the retrieved title and snippet pairs used ROUGE 1. The ROUGE 1 implementation also omitted stop words and stemmed all of the remaining content words using Porterstemmer. Many of the IDQ contain either "known" (e.g., "is known as") or "called" ("is called"). "Known" is on the stop word list, but "called" is not. There are additional comparison metrics that were run and are presented in Section 3.2.2 and whose results are explained in Section 4.2.2. Specifically, introducing WordNet helped improve the results of the comparison system. WordNet expanded the matching of synonyms for some words.

Figure 3.5 shows the ROUGE scores for the text strings from Example 6 that were presented in Figure F.8, except for the numbers that had zero scores. The first number in the line, in this case "151," corresponds to the number of the retrieved result being compared to the model_path, or original question. As detailed in Figure F.7, this means that 151, 152, and 153 are the first three results for answer option "cellular respiration" and number 162 corresponds to the twelfth retrieved result. The version of ROUGE used in the comparison is listed after the retrieved result number, followed by the average recall ("Average_R"), average precision ("Average_P"), or average F-score ("Average_F") and the numerical score with a confidence interval ("conf. int."). The lines in between the dotted line and the dashed line summarize all of the per question results in one line.

For number 151, the bigram scores were 0, revealing that while there was some individual word overlap, there were no overlapping two-word phrases. While the answer option "cellular respiration" was the correct answer to the question, retrieved results 154, 155, and 156 had no overlap with the text of the original question and as a result, are not shown in Figure 3.5. Also, questions 152 and 153 have no overlap when using the bigram comparison.

The precision score in Figure 3.5 corresponds to the original question having four words that overlap with the 17 content words in result D_12. The recall is 1.0 because all of the words in the question text occur in the retrieved title and snippet string. The F-score, calculated below, is also known as accuracy and is the metric used to judge the top results of the ROUGE module:

$$F = 2 \times \frac{precision \times recall}{precision + recall}.$$

Question 162 has the highest score due to the overlap of the terms "process," "releases," "energy," and "cell." Figure 3.6 presents the inverse definition text and text that ROUGE considers for comparison in three steps. Step one reiterates what is being compared: the text in the question stem to the retrieved results from "Define:" and each of the answer options. Step two shows what the returned text is for answer option D, "Cellular Respiration" and step three shows the actual text that is compared, the content words. In Figure 3.6, stop-words are shown crossed out, and the remaining words are in plain text.

The output files from ROUGE (Figure 3.5) were post-processed using a Python script and the top F-score result(s) for each question were sent to the appropriately numbered folder (e.g., in this case "82"). This top result could be unique or it could be

a tie with multiple retrieved results having the same F-score. In the case of a tie, there are several possible options:

- the ties are all from the correct answer option,
- the ties are from one or more answer options, including the correct answer option, and
- the ties are from neither of the above situations.

There were also four cases where there was no F-score; the ROUGE system shut down at some point in the processing. More information about this scenario and the scoring of tied results may be found in Section 4.1.1.

Next, the output files, like the one shown in part in Figure 3.5, are wrapped with XML tags for ease in the step of adding them to the Marklogic database. All of the information related to each question is in the directory for that question number, which also aided in structuring the XML database.

Post-Processing of ROUGE Results

The ROUGE module compared the original IDQ content words to those in 250 title and snippet strings. These 250 title and snippet strings were associated with the five answer options (A through E, as shown in Figure F.7). The ROUGE module measured the word overlap between the original IDQ and each of the 50 individual retrieved strings for each answer option. Then, the ROUGE module scored the results (as shown in Figure 3.5) and the highest F-score of this set was determined to be the best answer. This best answer corresponded to an answer option (A through E). Thus, in Figure 3.5, the best answer option was D_12, or the twelfth retrieved result of answer option D.

This best answer was then compared to the correct answer listed in the original question, as shown in Figure F.1. In this question, the best answer option D was also the correct answer. The ROUGE module could choose the correct answer, an incorrect answer, or have no answer. There could also be multiple top results if the top F-score was shared by two or three answers. These ties could contain correct answers, incorrect answers, or a mixture of both correct and incorrect answers. In the question example shown in Figure F.1, if there were two other D answer options that had the same F-score as D_12, the results would be a three-way correct tie for that question. If the correct answer was D, but there was a three-way tie for the top score between incorrect answers that corresponded to A, B, and C retrieved results, that question would be a three-way incorrect result.

Appendix G

Item Analysis Results and Cohort Movement Set1

The following pages present the bin movement of the top .15 most correlated students when they are graded on an exam comprised of the top .25 most correlated questions. Then only the most discriminating questions are used in an exam which is re-graded. Discriminating questions are those that have a positive Discriminating Power. Bin movement corresponds to students moving between performance groups based on the changed exam.

Set:	1
Total number of students:	1055
Total number of questions:	148
Percentage used of highest correlated students:	0.15
Percentage used of highest correlated questions:	0.25
Size of cohort:	158
Initial exam size:	37
New exam size:	24

After a summary of the bin movement, then follows the student-by-student results. The grades given to students corresponds with the percentage of the questions they answered correctly. Next each question in the top .25 most correlated question set is described including the text of the question, item analysis, and how the automated system judged each question.

In the examples of PeerWise questions that have been processed via Item Analysis, there is an additional column on the right hand side of each question analysis

titled "UFN" which stands for "usefulness." Usefulness of distractors is an elaboration by Mitkov et al. of the work on distractor effectiveness by Gronlund [65] [26]. A positive number in this column means that a distractor is "poor" since it attracts more high-performing students than lower-performing ones. A negative number is a "good" distractor that is chosen by more lower-performing students than higher performing ones. An "X" represents the correct answer. An "N" represents a not useful distractor since no students choose it. Finally, a "0" means that equal numbers of high and low performing students choose this option, so the distractor has no discriminating power. More information on this methodology may be found in Section 3.3 and an overview of the results may be found in Section 4.3.

Bin Movement Summary:

Low to middle:	4
Low to high:	0
Middle to low:	4
Middle to high:	5
High to low:	0
High to middle:	5
Total:	18(0.11)

Student ID	Initial Grade	New Grade	Initial Bin	New Bin
12838	0.16	0.15	low	low
11959	0.22	0.27	low	low
13030	0.22	0.23	low	low
11699	0.24	0.31	low	low
14020	0.27	0.31	low	low
12823	0.3	0.27	low	low
15067	0.32	0.27	low	low
11774	0.32	0.27	low	low
14322	0.32	0.31	low	low
13007	0.32	0.38	low	low
13516	0.32	0.31	low	low
11618	0.32	0.35	low	low
11767	0.32	0.23	low	low
12751	0.32	0.27	low	low
14135	0.35	0.46	low	low
14023	0.35	0.27	low	low
13895	0.35	0.35	low	low
12558	0.35	0.38	low	low
13074	0.35	0.35	low	low
13088	0.35	0.27	low	low
14076	0.35	0.27	low	low
12698	0.38	0.31	low	low
11924	0.38	0.42	low	low
11425	0.38	0.31	low	low
12926	0.38	0.46	low	middle
14114	0.38	0.42	low	low
12777	0.41	0.42	low	low
14064	0.41	0.38	low	low
13148	0.41	0.38	low	low
13829	0.41	0.5	low	middle
13096	0.41	0.42	low	low
13915	0.43	0.46	low	low
14320	0.43	0.54	low	middle
12584	0.43	0.46	low	low
12070	0.43	0.42	low	low
11812	0.43	0.46	low	low

10651	0.43	0.35	low	low
11764	0.46	0.46	low	low
12961	0.46	0.42	low	low
12193	0.46	0.46	low	low
13693	0.46	0.5	low	middle
13763	0.49	0.42	low	low
12994	0.49	0.46	low	low
13866	0.49	0.62	middle	middle
11579	0.49	0.54	middle	middle
13978	0.49	0.46	middle	low
10634	0.49	0.42	middle	low
13008	0.51	0.54	middle	middle
12375	0.51	0.58	middle	middle
11610	0.51	0.54	middle	middle
11496	0.51	0.5	middle	middle
10610	0.51	0.5	middle	middle
11510	0.51	0.58	middle	middle
13863	0.51	0.54	middle	middle
11570	0.51	0.5	middle	middle
12210	0.51	0.46	middle	low
13709	0.54	0.54	middle	middle
11155	0.54	0.58	middle	middle
10990	0.54	0.62	middle	middle
12957	0.54	0.58	middle	middle
13987	0.54	0.58	middle	middle
12929	0.54	0.54	middle	middle
12633	0.57	0.5	middle	middle
13526	0.57	0.62	middle	middle
13662	0.57	0.54	middle	middle
13421	0.57	0.42	middle	low
10619	0.57	0.5	middle	middle
11129	0.57	0.5	middle	middle
11411	0.57	0.58	middle	middle
11376	0.59	0.54	middle	middle
11126	0.59	0.58	middle	middle
12015	0.59	0.65	middle	middle
13826	0.59	0.69	middle	middle
13855	0.59	0.62	middle	middle
11422	0.59	0.58	middle	middle
14088	0.59	0.62	middle	middle
10623	0.62	0.58	middle	middle
10637	0.62	0.58	middle	middle
13508	0.62	0.69	middle	middle
14192	0.62	0.62	middle	middle
12933	0.62	0.62	middle	middle
12862	0.62	0.5	middle	middle

13627	0.65	0.62	middle	middle
11765	0.65	0.62	middle	middle
10944	0.65	0.65	middle	middle
13090	0.65	0.77	middle	high
12636	0.65	0.65	middle	middle
10993	0.65	0.58	middle	middle
12538	0.65	0.58	middle	middle
11645	0.65	0.62	middle	middle
12939	0.65	0.65	middle	middle
10854	0.65	0.54	middle	middle
11309	0.65	0.69	middle	middle
13018	0.65	0.65	middle	middle
11604	0.68	0.69	middle	middle
10629	0.68	0.62	middle	middle
10905	0.68	0.62	middle	middle
11584	0.68	0.62	middle	middle
13269	0.68	0.62	middle	middle
13740	0.68	0.62	middle	middle
13981	0.68	0.69	middle	middle
10680	0.68	0.73	middle	middle
11294	0.68	0.69	middle	middle
11628	0.7	0.65	middle	middle
12885	0.7	0.69	middle	middle
13827	0.7	0.62	middle	middle
12542	0.73	0.65	middle	middle
13006	0.73	0.77	middle	high
13100	0.73	0.69	middle	middle
11382	0.73	0.69	middle	middle
11307	0.73	0.73	middle	middle
12804	0.73	0.77	middle	high
11326	0.73	0.73	middle	middle
11357	0.73	0.77	middle	high
11329	0.76	0.81	middle	high
11476	0.76	0.73	middle	middle
11417	0.76	0.73	high	middle
10751	0.76	0.73	high	middle
12473	0.76	0.81	high	high
11995	0.76	0.73	high	high
13720	0.76	0.85	high	high
11373	0.76	0.73	high	high
12889	0.76	0.77	high	high
11148	0.76	0.73	high	middle
13106	0.76	0.73	high	high
10678	0.76	0.85	high	high
14121	0.76	0.81	high	high
10638	0.76	0.81	high	high

12912	0.76	0.73	high	middle
13787	0.76	0.65	high	middle
12756	0.76	0.81	high	high
12790	0.78	0.85	high	high
11607	0.78	0.73	high	high
11420	0.78	0.77	high	high
13668	0.78	0.77	high	high
11040	0.78	0.77	high	high
12786	0.81	0.81	high	high
13727	0.81	0.81	high	high
10836	0.81	0.81	high	high
14190	0.81	0.85	high	high
13144	0.81	0.81	high	high
13737	0.81	0.81	high	high
13626	0.84	0.85	high	high
13933	0.84	0.81	high	high
13937	0.84	0.88	high	high
11386	0.84	0.81	high	high
11150	0.84	0.85	high	high
14035	0.84	0.81	high	high
12176	0.86	0.85	high	high
10974	0.86	0.81	high	high
12934	0.86	0.88	high	high
10646	0.86	0.85	high	high
11132	0.86	0.92	high	high
12646	0.89	0.85	high	high
11478	0.89	0.88	high	high
11173	0.89	0.85	high	high
10642	0.92	0.96	high	high
12689	0.97	0.96	high	high

Question 29081 Average Human Rating: 3.1930 Average Human Difficulty: 0.6316

Which primary cell wall component exists in the crystalline microfibrillar phase?

A. Cellulose *correct answer*

B. Hemicellulose

C. Pectin

D. Protein

Letter	High	Middle	Low	Total	Diff	Automated System	
A	29	51	16	96	13	0.0151472	Correct
B	6	6	2	14	4	0.0333258	
C	7	6	1	14	6	0.0043314	
D	0	2	0	2	0	0.0	
E	0	0	0	0	0	0.0	
OMIT	0	8	24	32	0		
TOTAL	42	73	43	158	0		

Discriminating power: 3
Omission rate: 0.2
Question difficulty: 0.6076

Question 35490 Average Human Rating: 2.0635 Average Human Difficulty: 0.1429

What is the solution inside the central vacuole called?

A. Cell Sap *correct answer*
B. Cell Rap
C. Cell Hap
D. Cell Cype

Letter	High	Middle	Low	Total	Diff	Automated System	
A	41	58	18	117	23	0.0256394	Correct
B	0	0	0	0	0	0.003077	
C	1	0	0	1	1	0.00125	
D	0	1	1	2	-1	0.003205	
E	0	0	0	0	0	0.0	
OMIT	0	14	24	38	0		
TOTAL	42	73	43	158	0		

Discriminating power: 1
Omission rate: 0.24
Question difficulty: 0.7405

Question 35769 Average Human Rating: 3.1017 Average Human Difficulty: 0.5085

What is used in the Citric Acid Cycle to produce 2 ATP molecules?

A. Glucose
B. Pyruvate *correct answer*
C. Electron Transport Chain
D. Chemiosmosis
E. Oxidative phosphorilation

Letter	High	Middle	Low	Total	Diff	Automated System	
A	5	11	3	19	2	0.0042612	
B	34	35	14	83	20	0.0238406	Correct
C	2	4	0	6	2	0.0438152	
D	1	3	1	5	0	0.0298772	
E	0	3	1	4	-1	0.0332884	
OMIT	0	17	24	41	0		
TOTAL	42	73	43	158	0		

Discriminating power: 2
Omission rate: 0.26
Question difficulty: 0.5253

Question 27094 Average Human Rating: 3.6078 Average Human Difficulty: 0.9118

I am an organelle that can be found in animal cells. I can increase in number by splitting in two when I reach a certain size. I am bounded by a single membrane. My by-product and enzyme used to detoxify it are sequestered in the same space, away from other cellular components to prevent damaging them. What is my name?

A. Golgi Apparatus
B. Lysosome
C. Peroxisome *correct answer*
D. Mitochondria
E. Plasmodesmata

Letter	High	Middle	Low	Total	Diff	Automated System	
A	2	3	0	5	2	0.0414416	
B	15	19	11	45	4	0.0496438	
C	23	23	17	63	6	0.0382814	Correct
D	2	4	0	6	2	0.062921	
E	0	0	0	0	0	0.0370188	
OMIT	0	24	15	39	0		
TOTAL	42	73	43	158	0		

Discriminating power: 4
Omission rate: 0.25
Question difficulty: 0.3987

Question 32131 Average Human Rating: 3.1406 Average Human Difficulty: 0.3281

In which phase of meiosis do homologous chromosomes separate?

A. Anaphase II
B. Metaphase II
C. Anaphase I *correct answer*
D. Prophase II

E. Telophase I

Letter	High	Middle	Low	Total	Diff	Automated System	
A	3	11	2	16	1	0.1102208	
B	4	4	2	10	2	0.0673916	
C	34	51	24	109	10	0.0	Correct
D	1	1	0	2	1	0.0802086	
E	0	0	0	0	0	0.073093	
OMIT	0	6	15	21	0		
TOTAL	42	73	43	158	0		

Discriminating power: 4
Omission rate: 0.13
Question difficulty: 0.6899

Question 35480 Average Human Rating: 2.4211 Average Human Difficulty: 0.7018

What is the simplest form of starch?

A. Maltose

B. Amylopectin

C. Chloroplast

D. Amylose *correct answer*

E. Pectin

Letter	High	Middle	Low	Total	Diff	Automated System	
A	10	12	2	24	8	0.0404834	
B	3	7	1	11	2	0.031782	
C	0	2	0	2	0	0.0039046	
D	25	35	13	73	12	0.0357656	Correct
E	4	6	6	16	-2	0.0	
OMIT	0	11	21	32	0		
TOTAL	42	73	43	158	0		

Discriminating power: 2
Omission rate: 0.2
Question difficulty: 0.462

Question 30082 Average Human Rating: 2.8358 Average Human Difficulty: 0.4328

In the glucolysis stage of cellular respiration what is glucose converted to?

A. Glycogen

B. Citric Acid

C. Phosphorus

D. Pyruvate *correct answer*

E. Carbon Dioxide

Letter	High	Middle	Low	Total	Diff	Automated System	
A	4	11	3	18	1	0.016059	
B	0	1	0	1	0	0.0013794	
C	0	0	0	0	0	0.0	
D	37	52	20	109	17	0.004461	Correct
E	1	3	2	6	-1	0.0084346	
OMIT	0	6	18	24	0		
TOTAL	42	73	43	158	0		

Discriminating power: 1
Omission rate: 0.15
Question difficulty: 0.6899

Question 35229 Average Human Rating: 3.0517 Average Human Difficulty: 0.2414

Ribosome subunits are produced where?

A. Nucleolus *correct answer*

B. Nuclear Lamina

C. Nuclear Membrane

D. Nuclear Envelope

E. Rough Endoplasmic Reticulum

Letter	High	Middle	Low	Total	Diff	Automated System	
A	37	57	26	120	11	0.0043332	Correct
B	1	0	0	1	1	0.0	
C	0	1	0	1	0	0.0025	
D	0	2	1	3	-1	0.0	
E	3	2	2	7	1	0.02433	
OMIT	1	11	14	26	0		
TOTAL	42	73	43	158	0		

Discriminating power: 2
Omission rate: 0.16
Question difficulty: 0.7595

Question 30477 Average Human Rating: 3.1970 Average Human Difficulty: 0.5455

Where does the Krebs cycle (citric acid cycle) occur?

A. Intermembrane Space of Mitochondria

B. Cytosol

C. Mitochondrial Matrix *correct answer*

D. Stroma

E. Thylakoid

Letter	High	Middle	Low	Total	Diff	Automated System	
A	8	7	4	19	4	0.0016668	
B	3	3	1	7	2	0.0016	
C	26	53	21	100	5	0.0180358	Correct
D	2	5	1	8	1	0.0014284	
E	3	1	2	6	1	0.0055558	
OMIT	0	4	14	18	0		
TOTAL	42	73	43	158	0		

Discriminating power: 5
Omission rate: 0.11
Question difficulty: 0.6329

Question 35945 Average Human Rating: 3.4194 Average Human Difficulty: 0.6452

What is the disease associated with the deletion of a short arm of chromosome 5?

A. Klinefelter Syndrome
B. Lejeune Syndrome (Cri du chat) *correct answer*
C. Williams-Beuren Syndrome
D. Duchenne Muscular Dystrophy
E. Chronic Myeloid Leukemia

Letter	High	Middle	Low	Total	Diff	Automated System	
A	2	2	1	5	1	0.0530242	
B	26	39	20	85	6	0.0772464	Correct
C	2	7	5	14	-3	0.0313548	
D	8	3	5	16	3	0.0104594	
E	2	4	2	8	0	0.0090248	
OMIT	2	18	10	30	0		
TOTAL	42	73	43	158	0		

Discriminating power: 2
Omission rate: 0.19
Question difficulty: 0.538

Question 35735 Average Human Rating: 3.1897 Average Human Difficulty: 0.3966

Familial Down Syndrome is caused by?

A. Polyploidy
B. Translocation *correct answer*
C. Deletion
D. Inversion
E. Aneuploidy

Letter	High	Middle	Low	Total	Diff	Automated System	
A	3	5	2	10	1	0.0	
B	29	38	17	84	12	0.0136064	Correct
C	2	1	0	3	2	0.0114728	
D	0	1	1	2	-1	0.0	
E	7	15	5	27	2	0.0071146	
OMIT	1	13	18	32	0		
TOTAL	42	73	43	158	0		

Discriminating power: 3
Omission rate: 0.2
Question difficulty: 0.5616

Question 35241 Average Human Rating: 2.9701 Average Human Difficulty: 0.2388

Where are ribosomal subunits synthesized?

A. Nucleolus *correct answer*
B. Golgi Apparatus
C. Rough ER
D. Mitochondria
E. ECM

Letter	High	Middle	Low	Total	Diff	Automated System	
A	38	52	18	108	20	0.0043332	Correct
B	0	2	2	4	-2	0.0017392	
C	3	8	5	16	-2	0.0434778	
D	0	2	0	2	0	0.0038222	
E	1	1	1	3	0	0.0	
OMIT	0	8	17	25	0		
TOTAL	42	73	43	158	0		

Discriminating power: -1
Omission rate: 0.16
Question difficulty: 0.6835

Question 32127 Average Human Rating: 3.1316 Average Human Difficulty: 0.5395

In what phase does chromosome replication occur?

A. G2 Phase
B. Mitotic Phase
C. Prophase
D. G1 Phase
E. S Phase *correct answer*

Letter	High	Middle	Low	Total	Diff	Automated System	
A	9	5	4	18	5	0.0850532	
B	1	2	2	5	-1	0.0889784	
C	5	13	4	22	1	0.0378536	
D	4	8	1	13	3	0.098195	
E	23	40	17	80	6	0.1070478	Correct
OMIT	0	5	15	20	0		
TOTAL	42	73	43	158	0		

Discriminating power: 3
Omission rate: 0.13
Question difficulty: 0.5063

Question 35502 Average Human Rating: 3.1159 Average Human Difficulty: 0.2029

A cell maintains or changes shape via?

A. Gap Junctions
B. Cytoskeleton *correct answer*
C. Proteoglycans
D. Extracellular Matrix

Letter	High	Middle	Low	Total	Diff	Automated System	
A	0	0	0	0	0	0.0275562	
B	40	58	27	125	13	0.0616212	Correct
C	1	1	0	2	1	0.0177522	
D	1	4	0	5	1	0.0409144	
E	0	0	0	0	0	0.0	
OMIT	0	10	16	26	0		
TOTAL	42	73	43	158	0		

Discriminating power: 3
Omission rate: 0.16
Question difficulty: 0.7911

Question 33286 Average Human Rating: 3.5844 Average Human Difficulty: 0.5455

The gene in a fruit fly that controls eye color also controls wing span and body hair and this is an example of?

A. Polymorphism
B. Polygenism
C. Incomplete Dominance
D. Pleiotropy *correct answer*
E. Epistasis

Letter	High	Middle	Low	Total	Diff	Automated System	
A	8	8	4	20	4	0.0	
B	10	13	7	30	3	0.001081	
C	1	1	1	3	0	0.0111854	
D	21	34	18	73	3	0.0348356	Correct
E	2	4	6	12	-4	0.0304968	
OMIT	0	13	7	20	0		
TOTAL	42	73	43	158	0		

Discriminating power: 2
Omission rate: 0.13
Question difficulty: 0.462

Question 35681 Average Human Rating: 3.1791 Average Human Difficulty: 0.4925

A defect in the chloride ion transporter channel is responsible for what disease?

A. Type II Albinism
B. Cystic Fibrosis *correct answer*
C. Wilson's Disease
D. Epilepsy
E. Neurofibromatosis

Letter	High	Middle	Low	Total	Diff	Automated System	
A	1	1	0	2	1	0.0104756	
B	30	50	22	102	8	0.0241672	Correct
C	2	3	1	6	1	0.0797948	
D	5	5	1	11	4	0.0018182	
E	4	4	0	8	4	0.0140792	
OMIT	0	10	19	29	0		
TOTAL	42	73	43	158	0		

Discriminating power: 5
Omission rate: 0.18
Question difficulty: 0.6456

Question 35228 Average Human Rating: 2.9403 Average Human Difficulty: 0.2537

What is a symptom many aneuploids suffer from?

A. Infertility *correct answer*
B. Shortness
C. Poor Breast Development
D. Visual Deformities
E. Short Gestation Period

Letter	High	Middle	Low	Total	Diff	Automated System	
A	38	54	20	112	18	0.0059454	Correct
B	1	3	2	6	-1	0.0145124	
C	0	1	0	1	0	0.0041666	
D	2	6	7	15	-5	0.0015998	
E	0	0	0	0	0	0.0	
OMIT	1	9	14	24	0		
TOTAL	42	73	43	158	0		

Discriminating power: -1
Omission rate: 0.15
Question difficulty: 0.7089

Question 31325 Average Human Rating: 3.3514 Average Human Difficulty: 0.4054

A cell ingests droplets of extracellular fluid into vesicles, what would be the most accurate/specific term to describe this?

A. Phagocytosis
B. Endocytosis
C. Pinocytosis *correct answer*
D. Exocytosis
E. Osmosis

Letter	High	Middle	Low	Total	Diff	Automated System	
A	5	5	1	11	4	0.0558036	
B	5	12	4	21	1	0.0527138	
C	31	48	28	107	3	0.0727498	Correct
D	0	1	0	1	0	0.0493238	
E	1	2	0	3	1	0.0214402	
OMIT	0	5	10	15	0		
TOTAL	42	73	43	158	0		

Discriminating power: 4
Omission rate: 0.095
Question difficulty: 0.6772

Question 32237 Average Human Rating: 3.3108 Average Human Difficulty: 0.5000

At what stage(s) of meiosis can non disjunction occur?

A. Metaphase (I)
B. Anaphase (I and II) *correct answer*
C. Anaphase and Pro Metaphase (II)
D. Telephase (I and II)
E. Telephase (II)

Letter	High	Middle	Low	Total	Diff	Automated System	
A	4	7	5	16	-1	0.0513234	
B	34	49	30	113	4	0.0691068	Correct
C	2	5	2	9	0	0.0461916	
D	2	4	1	7	1	0.0265982	
E	0	1	1	2	-1	0.02427	
OMIT	0	7	4	11	0		
TOTAL	42	73	43	158	0		

Discriminating power: 0
Omission rate: 0.07
Question difficulty: 0.7152

Question 29083 Average Human Rating: 3.4571 Average Human Difficulty: 0.6286

What is found only in the secondary cell wall and not in the primary cell wall?

A. Cellulose

B. Hemicellulose

C. Lignin *correct answer*

D. Pectin

E. Extensin

Letter	High	Middle	Low	Total	Diff	Automated System	
A	0	2	0	2	0	0.018657	
B	3	2	2	7	1	0.0368496	
C	32	55	22	109	10	0.032482	Correct
D	3	2	1	6	2	0.0104626	
E	4	7	2	13	2	0.045657	
OMIT	0	5	16	21	0		
TOTAL	42	73	43	158	0		

Discriminating power: 4
Omission rate: 0.13
Question difficulty: 0.6899

Question 35440 Average Human Rating: 3.2133 Average Human Difficulty: 0.7600

When a specific substance is ingested into a cell from a low concentration, describes?

A. Active Transport

B. Pinocytosis

C. Phagocytosis

D. Receptor Mediated Endocytosis *correct answer*

E. Facilitated Diffusion

Letter	High	Middle	Low	Total	Diff	Automated System	
A	14	21	5	40	9	0.0587048	
B	5	4	4	13	1	0.0442784	
C	7	9	6	22	1	0.057833	
D	15	29	13	57	2	0.0418652	Correct
E	1	2	1	4	0	0.0471664	
OMIT	0	8	14	22	0		
TOTAL	42	73	43	158	0		

Discriminating power: 4
Omission rate: 0.14
Question difficulty: 0.3608

Question 35230 Average Human Rating: 3.2727 Average Human Difficulty: 0.3636

Facilitated Diffusion involves the moving of specific substances down their concentration gradient. What does this actively involve?

A. ATP

B. Peripheral Proteins

C. Channel Proteins *correct answer*

D. Pumps

Letter	High	Middle	Low	Total	Diff	Automated System	
A	3	3	1	7	2	0.0038328	
B	0	0	1	1	-1	0.0125354	
C	35	61	31	127	4	0.0148422	Correct
D	3	2	3	8	0	0.0115712	
E	0	0	0	0	0	0.0	
OMIT	1	7	7	15	0		
TOTAL	42	73	43	158	0		

Discriminating power: 1
Omission rate: 0.095
Question difficulty: 0.8038

Question 31761 Average Human Rating: 3.4945 Average Human Difficulty: 0.9780

A sperm cell about to undergo meiosis II is called a?

A. Spermatagonia

B. Secondary Spermatocyte *correct answer*

C. Oogonia

D. Primary Spermatocyte

E. Spermatid

Letter	High	Middle	Low	Total	Diff	Automated System	
A	13	7	4	24	9	0.0640448	
B	17	36	14	67	3	0.07034	Correct
C	1	3	2	6	-1	0.0246144	
D	7	12	13	32	-6	0.0529098	
E	4	10	3	17	1	0.031517	
OMIT	0	5	7	12	0		
TOTAL	42	73	43	158	0		

Discriminating power: 1
Omission rate: 0.076
Question difficulty: 0.4241

Question 34932 Average Human Rating: 3.4054 Average Human Difficulty: 0.7838

Extensin is synthesized in the?

A. Golgi Apparatus
B. Mitochondria
C. Smooth Endoplasmic Reticulum
D. Rough Endoplasmic Reticulum *correct answer*
E. Nucleus

Letter	High	Middle	Low	Total	Diff	Automated System	
A	6	19	10	35	-4	0.0	
B	1	3	1	5	0	0.0	
C	5	6	5	16	0	0.0020002	
D	28	41	20	89	8	0.0063896	Correct
E	1	2	2	5	-1	0.0	
OMIT	1	2	5	8	0		
TOTAL	42	73	43	158	0		

Discriminating power: -1
Omission rate: 0.051
Question difficulty: 0.5633

Question 35352 Average Human Rating: 3.4605 Average Human Difficulty: 0.6184

Microtubules attach to what structure of Homologous Chromosomes at Metaphase 1 during Meiosis?

A. Centromere
B. Kinetochore *correct answer*
C. Tetrad
D. Chromatin

Letter	High	Middle	Low	Total	Diff	Automated System	
A	15	18	10	43	5	0.051567	
B	23	45	24	92	-1	0.0587048	Correct
C	1	2	1	4	0	0.010484	
D	1	1	0	2	1	0.0294984	
E	0	0	0	0	0	0.0	
OMIT	2	7	8	17	0		
TOTAL	42	73	43	158	0		

Discriminating power: 1
Omission rate: 0.11
Question difficulty: 0.5823

Question 34818 Average Human Rating: 3.4500 Average Human Difficulty: 0.7875

In the mitochondria, high energy electrons are extracted from the oxidation of?

A. Water Molecules

B. ATP

C. Light Energy Photons

D. Organic Molecules *correct answer*

Letter	High	Middle	Low	Total	Diff	Automated System	
A	8	7	4	19	4	0.0105646	
B	13	20	10	43	3	0.0167534	
C	1	2	3	6	-2	0.079477	
D	19	41	22	82	-3	0.0071288	Correct
E	0	0	0	0	0	0.0	
OMIT	1	3	4	8	0		
TOTAL	42	73	43	158	0		

Discriminating power: 0
Omission rate: 0.051
Question difficulty: 0.519

Question 35104 Average Human Rating: 3.2273 Average Human Difficulty: 0.7841

Failure of the CFTR (Cl-) channel results in which disease?

A. Cystic Fibrosis *correct answer*

B. Epilepsy

C. Wilson's Disease

D. Albinism (Type 2)

E. Neurofibromatosis

Letter	High	Middle	Low	Total	Diff	Automated System	
A	35	55	30	120	5	0.02867	Correct
B	3	7	2	12	1	0.0032382	
C	2	4	4	10	-2	0.0841026	
D	1	2	1	4	0	0.016393	
E	1	1	1	3	0	0.0147314	
OMIT	0	4	5	9	0		
TOTAL	42	73	43	158	0		

Discriminating power: 1
Omission rate: 0.057
Question difficulty: 0.7595

Question 35409 Average Human Rating: 3.4494 Average Human Difficulty: 0.5618

In Klinefelter syndrome, individuals are phenotypically male, but they have reduced sperm production and may have some breast development in adolescence. The cells of Klinefelter individuals have two X chromosomes and one Y (they are XXY instead of XY). This occurs because of what meotic error?

A. Translocation
B. Polyploidy
C. Aneuploidy *correct answer*
D. Duplication
E. Monosomy

Letter	High	Middle	Low	Total	Diff	Automated System	
A	4	1	1	6	3	0.033645	
B	7	13	4	24	3	0.028196	
C	29	48	24	101	5	0.035528	Correct
D	2	2	4	8	-2	0.0083764	
E	0	1	2	3	-2	0.0300022	
OMIT	0	8	8	16	0		
TOTAL	42	73	43	158	0		

Discriminating power: 1
Omission rate: 0.1
Question difficulty: 0.6392

Question 34960 Average Human Rating: 3.2941 Average Human Difficulty: 0.9608

Which of the following cellular structures does Hutchinson-Gilford Progeria Syndrome directly affect?

A. Mitochondria
B. Lysosomes
C. Peroxisomes

D. Channel Proteins

E. Nuclear Lamina *correct answer*

Letter	High	Middle	Low	Total	Diff	Automated System	
A	2	3	3	8	-1	0.0241242	
B	1	3	4	8	-3	0.0115626	
C	6	5	4	15	2	0.0153946	
D	4	5	8	17	-4	0.0076198	
E	28	52	21	101	7	0.0225154	Correct
OMIT	1	5	3	9	0		
TOTAL	42	73	43	158	0		

Discriminating power: -1
Omission rate: 0.057
Question difficulty: 0.6392

Question 35589 Average Human Rating: 3.6387 Average Human Difficulty: 0.5294

Micro-tubles are involved with which kind of cell motility?

A. Cytoplasmic Streaming

B. Pseudopodia

C. Locomotion (Cilia and Flagella) *correct answer*

D. Chromosomal Division (Spindle Fibres)

E. Organelle Movement

Letter	High	Middle	Low	Total	Diff	Automated System	
A	4	4	4	12	0	0.021714	
B	0	1	0	1	0	0.013678	
C	28	43	19	90	9	0.0445986	Correct
D	2	5	5	12	-3	0.0483194	
E	8	14	6	28	2	0.0572704	
OMIT	0	6	9	15	0		
TOTAL	42	73	43	158	0		

Discriminating power: 1
Omission rate: 0.095
Question difficulty: 0.5696

Question 34905 Average Human Rating: 3.5702 Average Human Difficulty: 0.8333

Which substance exists in the crystalline micro-fibrillar phase?

A. Pectin

B. Extensin

C. Cellulose *correct answer*

D. Lignin

E. Hemi-Cellulose

Letter	High	Middle	Low	Total	Diff	Automated System	
A	3	4	6	13	-3	0.0	
B	0	6	3	9	-3	0.0	
C	33	45	21	99	12	0.0030284	Correct
D	3	8	4	15	-1	0.0095274	
E	3	8	5	16	-2	0.0034846	
OMIT	0	2	4	6	0		
TOTAL	42	73	43	158	0		

Discriminating power: -3
Omission rate: 0.038
Question difficulty: 0.6266

Appendix H

Item Analysis Results and Cohort Movement Set2

The following pages present the bin movement of the top .15 most correlated students when they are graded on an exam comprised of the top .25 most correlated questions. Then only the most discriminating questions are used in an exam which is re-graded. Discriminating questions are those that have a positive Discriminating Power. Bin movement corresponds to students moving between performance groups based on the changed exam.

After a summary of the bin movement, then follows the student-by-student results. The grades given to students corresponds with the percentage of the questions they answered correctly. Next each question in the top .25 most correlated question set is described including the text of the question, item analysis, and how the automated system judged each question.

In the examples of PeerWise questions that have been processed via Item Analysis, there is an additional column on the right hand side of each question analysis titled "UFN" which stands for "usefulness." Usefulness of distractors is an elaboration by Mitkov et al. of the work on distractor effectiveness by Gronlund [65] [26]. A positive number in this column means that a distractor is "poor" since it attracts more high-performing students than lower-performing ones. A negative number is a "good" distractor that is chosen by more lower-performing students than higher performing ones. An "X" represents the correct answer. An "N" represents a not useful distractor since no students choose it. Finally, a "0" means that equal numbers of high and low performing students choose this option, so the distractor has no discriminating power. More information on this methodology may be found in Section 3.3 and an overview

of the results may be found in Section 4.3.

Bin Movement Summary:

Low to middle:	5
Low to high:	0
Middle to low:	5
Middle to high:	8
High to low:	0
High to middle:	8
Total:	26(0.2)

Student ID	Initial Grade	New Grade	Initial Bin	New Bin
15889	0.18	0.19	low	low
14064	0.27	0.33	low	low
12957	0.27	0.33	low	low
13609	0.33	0.38	low	low
13005	0.36	0.38	low	low
12935	0.36	0.38	low	low
10994	0.36	0.33	low	low
13100	0.36	0.38	low	low
11126	0.36	0.38	low	low
12769	0.39	0.38	low	low
15067	0.42	0.29	low	low
13906	0.45	0.38	low	low
11512	0.45	0.48	low	low
13088	0.45	0.43	low	low
12910	0.45	0.48	low	low
13516	0.45	0.57	low	low
10686	0.45	0.43	low	low
11155	0.48	0.48	low	low
11610	0.48	0.43	low	low
13845	0.48	0.48	low	low
12796	0.52	0.52	low	low
11333	0.52	0.52	low	low
12839	0.52	0.48	low	low
10646	0.55	0.57	low	low
15254	0.55	0.52	low	low
13030	0.55	0.52	low	low
12706	0.58	0.57	low	middle
13106	0.58	0.62	low	middle
11765	0.58	0.57	low	low
13987	0.58	0.52	low	low
10836	0.58	0.57	low	low
15472	0.58	0.52	low	low
12840	0.58	0.57	low	low
12210	0.58	0.62	low	middle
15473	0.58	0.57	low	middle
13978	0.58	0.67	low	middle

14135	0.58	0.67	middle	middle
13918	0.61	0.57	middle	middle
12070	0.61	0.76	middle	middle
15846	0.61	0.67	middle	middle
12838	0.61	0.52	middle	low
13866	0.61	0.57	middle	middle
12962	0.61	0.62	middle	middle
11376	0.61	0.62	middle	middle
15347	0.61	0.48	middle	low
15348	0.61	0.57	middle	middle
11735	0.64	0.71	middle	middle
13008	0.64	0.62	middle	middle
12961	0.64	0.57	middle	middle
13483	0.64	0.57	middle	middle
14365	0.64	0.76	middle	middle
11420	0.64	0.67	middle	middle
12804	0.64	0.52	middle	low
13139	0.64	0.57	middle	middle
11995	0.64	0.57	middle	middle
13081	0.67	0.62	middle	middle
13800	0.67	0.67	middle	middle
12212	0.67	0.57	middle	low
13827	0.67	0.62	middle	middle
15899	0.67	0.57	middle	middle
12753	0.67	0.62	middle	middle
14156	0.67	0.57	middle	low
11451	0.67	0.62	middle	middle
12636	0.67	0.57	middle	middle
12633	0.7	0.57	middle	middle
12912	0.7	0.71	middle	middle
11382	0.7	0.71	middle	middle
14185	0.7	0.71	middle	middle
11309	0.7	0.67	middle	middle
15200	0.7	0.67	middle	middle
12698	0.7	0.62	middle	middle
11476	0.7	0.71	middle	middle
10678	0.73	0.62	middle	middle
12077	0.73	0.76	middle	high
10623	0.73	0.71	middle	middle
10916	0.73	0.71	middle	middle
13114	0.73	0.76	middle	middle
10751	0.73	0.76	middle	middle
15255	0.73	0.67	middle	middle
10648	0.76	0.62	middle	middle
13627	0.76	0.71	middle	middle
12786	0.76	0.71	middle	middle

14190	0.76	0.76	middle	middle
14121	0.76	0.71	middle	middle
12756	0.76	0.67	middle	middle
11132	0.76	0.81	middle	high
10680	0.76	0.67	middle	middle
11645	0.76	0.86	middle	high
10651	0.76	0.71	middle	middle
11604	0.76	0.81	middle	high
11307	0.76	0.71	middle	middle
11148	0.79	0.76	middle	middle
10993	0.79	0.76	middle	middle
13144	0.79	0.76	middle	middle
12584	0.79	0.81	middle	high
15342	0.79	0.86	middle	high
14157	0.79	0.81	middle	high
11699	0.79	0.81	middle	high
14361	0.79	0.67	high	middle
12645	0.79	0.81	high	high
13074	0.79	0.76	high	middle
12375	0.79	0.71	high	middle
11837	0.79	0.71	high	middle
12176	0.79	0.76	high	middle
11386	0.82	0.81	high	high
10637	0.82	0.81	high	high
10629	0.82	0.81	high	high
11478	0.82	0.71	high	middle
12145	0.82	0.86	high	high
10974	0.82	0.76	high	middle
15410	0.82	0.76	high	middle
11129	0.85	0.86	high	high
10905	0.85	0.95	high	high
13102	0.85	0.86	high	high
12790	0.85	0.86	high	high
13626	0.85	0.81	high	high
12646	0.85	0.86	high	high
12722	0.88	0.81	high	high
11150	0.88	0.9	high	high
11628	0.88	0.81	high	high
10642	0.88	0.9	high	high
13787	0.88	0.9	high	high
13951	0.91	0.9	high	high
12885	0.91	1	high	high
11173	0.91	0.9	high	high
14035	0.91	0.95	high	high
11722	0.91	0.9	high	high
13006	0.91	0.86	high	high

12933	0.94	0.9	high	high
13269	0.94	0.9	high	high
12542	0.94	0.95	high	high
12689	0.97	0.95	high	high
12667	0.97	1	high	high

Question 41323 Average Human Rating: 3.5200 Average Human Difficulty: 0.4091

I am associated with blood vessels, I stretch over sulci and gyri to form a smooth layer, and I have extensions that protrude into venous sinuses, providing a path for CSF to enter the blood. What am I?

A. Dura Mater

B. Tentorium Cerebelli

C. Falx Cerebri

D. Arachnoid Mater *correct answer*

E. Pia Mater

Letter	High	Middle	Low	Total	Diff	Automated System	
A	2	3	4	9	-2	0.0161246	
B	1	1	3	5	-2	0.0262154	
C	2	3	1	6	1	0.012282	
D	22	35	13	70	9	0.0162682	Correct
E	8	16	5	29	3	0.0090076	
OMIT	0	4	10	14	0		
TOTAL	35	62	36	133	0		

Discriminating power: 1
Omission rate: 0.11
Question difficulty: 0.5263

Question 41279 Average Human Rating: 2.8333 Average Human Difficulty: 0.3056

Which inhibitory transmitter causes hyperpolarization?

A. Noradrenaline

B. GABA *correct answer*

C. Acetylcholine

D. Norepinephrine

E. Glutamate

Letter	High	Middle	Low	Total	Diff	Automated System	
A	1	0	0	1	1	0.0	
B	30	53	25	108	5	0.0192588	Correct
C	2	2	2	6	0	0.0017392	
D	0	3	1	4	-1	0.0021052	
E	2	1	3	6	-1	0.0	
OMIT	0	3	5	8	0		
TOTAL	35	62	36	133	0		

Discriminating power: 0
Omission rate: 0.06
Question difficulty: 0.812

Question 41443 Average Human Rating: 3.5600 Average Human Difficulty: 0.5600

Which hormone secretion pattern is directly affected from jet lag?

A. Cortisol *correct answer*
B. Insulin
C. Thyroid Hormone
D. Adrenaline
E. Calcitonin

Letter	High	Middle	Low	Total	Diff	Automated System	
A	24	40	17	81	7	0.016633	Correct
B	1	1	0	2	1	0.0315048	
C	9	12	3	24	6	0.06804	
D	0	2	1	3	-1	0.0095742	
E	1	0	2	3	-1	0.0388412	
OMIT	0	7	13	20	0		
TOTAL	35	62	36	133	0		

Discriminating power: 1
Omission rate: 0.15
Question difficulty: 0.609

Question 40328 Average Human Rating: 3.0698 Average Human Difficulty: 0.3953

Which describes the connective tissue that surrounds a single nerve fibre?

A. Fasicle
B. Perineurium
C. Epineurium
D. Endoneurium *correct answer*

Letter	High	Middle	Low	Total	Diff	Automated System	
A	4	1	1	6	3	0.0143652	
B	4	3	3	10	1	0.0825952	
C	5	11	2	18	3	0.0772496	
D	22	46	23	91	-1	0.0705898	Correct
E	0	0	0	0	0	0.0	
OMIT	0	1	7	8	0		
TOTAL	35	62	36	133	0		

Discriminating power: 2
Omission rate: 0.06
Question difficulty: 0.6842

Question 43999 Average Human Rating: 3.2115 Average Human Difficulty: 0.5000

The injection of antibodies from an immunized animal into the non-immune patient, such as antivenom, is a type of immunization called?

A. Active Natural

B. Passive Natural

C. Active Artificial

D. Passive Artificial *correct answer*

Letter	High	Middle	Low	Total	Diff	Automated System	
A	1	0	2	3	-1	0.0322582	
B	7	5	3	15	4	0.0459386	
C	5	2	5	12	0	0.0150538	
D	22	43	16	81	6	0.025998	Correct
E	0	0	0	0	0	0.0	
OMIT	0	12	10	22	0		
TOTAL	35	62	36	133	0		

Discriminating power: 1
Omission rate: 0.17
Question difficulty: 0.609

Question 44001 Average Human Rating: 3.4694 Average Human Difficulty: 0.5306

What glia cell is responsible for inhibiting snake venom from entering the central nervous system from the blood stream?

A. Schwann cells

B. Ependymal cells

C. Microglia

D. Astrocytes *correct answer*

E. Oligodendrocytes

Letter	High	Middle	Low	Total	Diff	Automated System	
A	2	2	3	7	-1	0.0869202	
B	3	3	3	9	0	0.0868452	
C	5	7	2	14	3	0.0493588	
D	24	36	16	76	8	0.0311876	Correct
E	1	1	2	4	-1	0.0577894	
OMIT	0	13	10	23	0		
TOTAL	35	62	36	133	0		

Discriminating power: 0
Omission rate: 0.17
Question difficulty: 0.481

Question 41332 Average Human Rating: 3.3448 Average Human Difficulty: 0.9138

Which primary hormone is responsible for growth in a one-year-old child?

A. Growth Hormone

B. Cortisol

C. Insulin

D. Thyroid Hormone *correct answer*

E. IGF-1

Letter	High	Middle	Low	Total	Diff	Automated System	
A	18	28	9	55	9	0.1605618	
B	1	0	0	1	1	0.0206506	
C	1	0	0	1	1	0.0323486	
D	7	16	4	27	3	0.0820942	Correct
E	8	13	10	31	-2	0.0355844	
OMIT	0	5	13	18	0		
TOTAL	35	62	36	133	0		

Discriminating power: 3
Omission rate: 0.14
Question difficulty: 0.203

Question 41084 Average Human Rating: 3.3077 Average Human Difficulty: 0.3077

Cerebrospinal Fluid is generated by which part of the brain?

A. Corpus Callosum

B. Diencephalon

C. Choroid Plexus *correct answer*

D. Basal Ganglia

E. Endocrine System

Letter	High	Middle	Low	Total	Diff	Automated System	
A	3	4	2	9	1	0.0155786	
B	0	1	0	1	0	0.034723	
C	30	54	27	111	3	0.0334834	Correct
D	2	1	1	4	1	0.025868	
E	0	0	0	0	0	0.009327	
OMIT	0	2	6	8	0		
TOTAL	35	62	36	133	0		

Discriminating power: 3
Omission rate: 0.06
Question difficulty: 0.8346

Question 41366 Average Human Rating: 3.2444 Average Human Difficulty: 0.4889

Which condition can result from a hypersecretion (too much) of cortisol from the adrenal cortex?

A. Addison's Disease

B. Grave's Disease

C. Cretinism

D. Cushing Syndrome *correct answer*

Letter	High	Middle	Low	Total	Diff	Automated System	
A	3	9	4	16	-1	0.07117	
B	6	12	7	25	-1	0.0048342	
C	4	4	3	11	1	0.0316114	
D	22	35	15	72	7	0.0441618	Correct
E	0	0	0	0	0	0.0	
OMIT	0	2	7	9	0		
TOTAL	35	62	36	133	0		

Discriminating power: 0
Omission rate: 0.068
Question difficulty: 0.5414

Question 40524 Average Human Rating: 2.8696 Average Human Difficulty: 0.4130

What neurotransmitter is contained in the synaptic vesicles in neuromuscular junctions?

A. Acetyl Choline *correct answer*

B. GABA (Gamma Amino Butyric Acid)

C. Ca2+ ions

D. Noradrenaline

E. Glutamate

Letter	High	Middle	Low	Total	Diff	Automated System	
A	30	51	24	105	6	0.0211662	Correct
B	2	2	0	4	2	0.0318998	
C	2	5	6	13	-4	0.0028606	
D	1	2	1	4	0	0.0129384	
E	0	1	0	1	0	0.0038914	
OMIT	0	1	5	6	0		
TOTAL	35	62	36	133	0		

Discriminating power: 1
Omission rate: 0.045
Question difficulty: 0.7895

Question 41481 Average Human Rating: 3.1724 Average Human Difficulty: 0.6379

Which of the bones of the cranium is associated with the nasal cavity?

A. Frontal

B. Parietal

C. Occipital

D. Sphenoid

E. Ethmoid *correct answer*

Letter	High	Middle	Low	Total	Diff	Automated System	
A	6	8	4	18	2	0.0126918	
B	0	4	0	4	0	0.0366866	
C	0	0	0	0	0	0.0286016	
D	13	15	5	33	8	0.0640646	
E	16	29	17	62	-1	0.0493792	Correct
OMIT	0	6	10	16	0		
TOTAL	35	62	36	133	0		

Discriminating power: 1
Omission rate: 0.12
Question difficulty: 0.4662

Question 41329 Average Human Rating: 3.4103 Average Human Difficulty: 0.3590

What type of glia cell engulfs and destroys micro-organisms and debris?

A. Astrocytes

B. Microglia *correct answer*

C. Ependymal cells

D. Oligodendrocytes

E. Schwann cells

Letter	High	Middle	Low	Total	Diff	Automated System	
A	2	1	0	3	2	0.0441186	
B	33	54	24	111	9	0.0300684	Correct
C	0	2	4	6	-4	0.0863002	
D	0	1	2	3	-2	0.0438396	
E	0	0	0	0	0	0.0856166	
OMIT	0	4	6	10	0		
TOTAL	35	62	36	133	0		

Discriminating power: 0
Omission rate: 0.075
Question difficulty: 0.8346

Question 31221 Average Human Rating: 3.5488 Average Human Difficulty: 0.4146

Articular cartilage allows for frictionless movement between bones.
In which joint can articular cartilage be found?

A. Cartilaginous Joint

B. Fibrous Joint

C. Synovial Joint *correct answer*

D. Bony Congruence

Letter	High	Middle	Low	Total	Diff	Automated System	
A	1	1	1	3	0	0.1226036	
B	0	1	1	2	-1	0.1083344	
C	31	51	21	103	10	0.0867412	Correct
D	3	0	0	3	3	0.0275524	
E	0	0	0	0	0	0.0	
OMIT	0	9	13	22	0		
TOTAL	35	62	36	133	0		

Discriminating power: 1
Omission rate: 0.17
Question difficulty: 0.7744

Question 40435 Average Human Rating: 2.9111 Average Human Difficulty: 0.3333

Which hormone affects your BMR?

A. Calcitonin

B. Thyroid Hormone *correct answer*

C. PTH

D. Cortisol

Letter	High	Middle	Low	Total	Diff	Automated System	
A	5	4	2	11	3	0.0560442	
B	24	44	23	91	1	0.0880502	Correct
C	2	4	0	6	2	0.036287	
D	4	8	7	19	-3	0.0102242	
E	0	0	0	0	0	0.0	
OMIT	0	2	4	6	0		
TOTAL	35	62	36	133	0		

Discriminating power: 2
Omission rate: 0.045
Question difficulty: 0.6842

Question 40329 Average Human Rating: 2.9167 Average Human Difficulty: 0.3125

What is the sulcus/fissure called which divides the brain frontal from parietal?

A. Lateral Fissure

B. Parietoccipital Sulcus

C. Longitudinal Fissure

D. Central Sulcus *correct answer*

Letter	High	Middle	Low	Total	Diff	Automated System	
A	8	11	7	26	1	0.138894	
B	2	3	5	10	-3	0.0429212	
C	2	1	1	4	1	0.116559	
D	23	46	17	86	6	0.1415858	Correct
E	0	0	0	0	0	0.0	
OMIT	0	1	6	7	0		
TOTAL	35	62	36	133	0		

Discriminating power: 2
Omission rate: 0.053
Question difficulty: 0.6466

Question 29090 Average Human Rating: 3.6495 Average Human Difficulty: 0.2990

What bone on the appendicular skeleton is inferior to the humerus, proximal to the metacarpals and medial of the radius?

A. Coccyx

B. Fibula

C. Ulna *correct answer*

D. Humerus

E. Scapula

Letter	High	Middle	Low	Total	Diff	Automated System	
A	1	0	0	1	1	0.0141686	
B	0	1	2	3	-2	0.0290068	
C	34	54	20	108	14	0.0437344	Correct
D	0	0	0	0	0	0.093629	
E	0	0	0	0	0	0.0186022	
OMIT	0	7	14	21	0		
TOTAL	35	62	36	133	0		

Discriminating power: 1
Omission rate: 0.16
Question difficulty: 0.812

Question 40327 Average Human Rating: 2.7045 Average Human Difficulty: 0.2273

What types of muscle get fatigue?

A. Cardiac
B. Smooth
C. Skeletal *correct answer*
D. Rough
E. Long

Letter	High	Middle	Low	Total	Diff	Automated System	
A	0	0	0	0	0	0.0097592	
B	1	2	0	3	1	0.0068	
C	31	55	27	113	4	0.038075	Correct
D	1	0	0	1	1	0.0	
E	2	4	3	9	-1	0.0	
OMIT	0	1	6	7	0		
TOTAL	35	62	36	133	0		

Discriminating power: 2
Omission rate: 0.053
Question difficulty: 0.8496

Question 31976 Average Human Rating: 3.2297 Average Human Difficulty: 0.4730

Hyposecretion of insulin is associated with what disease?

A. Type 1 Hypoimmuno Disease
B. Type 2 Hyperglycemia
C. Type 1 Diabetes Mellitus *correct answer*
D. Type 2 Diabetes Mellitus
E. Cancer

Letter	High	Middle	Low	Total	Diff	Automated System	
A	1	1	1	3	0	0.0032664	
B	3	3	1	7	2	0.0275722	
C	18	39	16	73	2	0.0508464	Correct
D	12	8	3	23	9	0.0439516	
E	1	0	0	1	1	0.0141994	
OMIT	0	11	15	26	0		
TOTAL	35	62	36	133	0		

Discriminating power: 4
Omission rate: 0.2
Question difficulty: 0.5488

Question 40243 Average Human Rating: 3.4706 Average Human Difficulty: 0.4118

I am a layer of the meninges, which are the inner coverings of the brain and spinal cord. I am composed of dense fibrous connective tissue, and I am composed of two layers, which have a space between them that is filled with venous blood. What layer of the meninges am I?

A. Pia Mater

B. Choroid Plexus

C. Dura Mater *correct answer*

D. Arachnoid Mater

E. Periosteum

Letter	High	Middle	Low	Total	Diff	Automated System	
A	4	1	5	10	-1	0.0545448	
B	0	0	0	0	0	0.0326632	
C	25	49	21	95	4	0.065995	Correct
D	6	10	4	20	2	0.0904644	
E	0	0	0	0	0	0.0917294	
OMIT	0	2	6	8	0		
TOTAL	35	62	36	133	0		

Discriminating power: 1
Omission rate: 0.06
Question difficulty: 0.7143

Question 40439 Average Human Rating: 3.1042 Average Human Difficulty: 0.4167

The production of which hormone requires iodine?

A. Glucagon

B. Oxytocin

C. PTH

D. Prolactin

E. Thyroid Hormone *correct answer*

Letter	High	Middle	Low	Total	Diff	Automated System	
A	0	2	1	3	-1	0.0483924	
B	3	1	0	4	3	0.0437792	
C	1	4	0	5	1	0.0361802	
D	0	1	2	3	-2	0.0625662	
E	31	52	29	112	2	0.0879246	Correct
OMIT	0	2	4	6	0		
TOTAL	35	62	36	133	0		

Discriminating power: 1
Omission rate: 0.045
Question difficulty: 0.8421

Question 41416 Average Human Rating: 3.5902 Average Human Difficulty: 0.3443

Where in the brain is the primary visual cortex located?

A. Temporal Lobe

B. Cerrebellum

C. Precentral Gyrus

D. Postcentral Gyrus

E. Occipital Lobe *correct answer*

Letter	High	Middle	Low	Total	Diff	Automated System	
A	3	2	1	6	2	0.0338344	
B	2	1	0	3	2	0.0166912	
C	1	0	1	2	0	0.03623	
D	0	5	0	5	0	0.0293812	
E	29	51	29	109	0	0.0365114	Correct
OMIT	0	3	5	8	0		
TOTAL	35	62	36	133	0		

Discriminating power: 2
Omission rate: 0.06
Question difficulty: 0.8195

Question 40361 Average Human Rating: 3.3750 Average Human Difficulty: 0.5000

Which hormone is most affected by the lack of sleep?

A. Cortisol *correct answer*

B. PTH

C. ADH

D. Thyroid Hormone

Letter	High	Middle	Low	Total	Diff	Automated System	
A	25	48	21	94	4	0.0405404	Correct
B	1	4	3	8	-2	0.0366704	
C	3	5	5	13	-2	0.0325538	
D	6	3	1	10	5	0.0745484	
E	0	0	0	0	0	0.0	
OMIT	0	2	6	8	0		
TOTAL	35	62	36	133	0		

Discriminating power: 0
Omission rate: 0.06
Question difficulty: 0.7068

Question 41280 Average Human Rating: 3.5349 Average Human Difficulty: 0.5349

The afferent pathway for posture consists of three neurons in relay, is remembered by 'up and across' and is called the?

A. Lateral Spinothalamic Pathway
B. Medial Lemniscal Pathway *correct answer*
C. Final Common Pathway
D. Anterior Thyrominiscal Pathway
E. Corticospinal pathway (pyrimidal tract)

Letter	High	Middle	Low	Total	Diff	Automated System	
A	3	5	10	18	-7	0.0841164	
B	28	50	19	97	9	0.0867422	Correct
C	0	0	0	0	0	0.0809798	
D	0	0	0	0	0	0.0	
E	4	5	0	9	4	0.0	
OMIT	0	2	7	9	0		
TOTAL	35	62	36	133	0		

Discriminating power: 1
Omission rate: 0.068
Question difficulty: 0.7293

Question 40973 Average Human Rating: 3.3750 Average Human Difficulty: 0.6042

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

Letter	High	Middle	Low	Total	Diff	Automated System	
A	3	7	6	16	-3	0.0413738	
B	25	41	16	82	9	0.0016	Correct
C	2	2	1	5	1	0.0317248	
D	0	3	3	6	-3	0.0014814	
E	5	7	6	18	-1	0.0095852	
OMIT	0	2	4	6	0		
TOTAL	35	62	36	133	0		

Discriminating power: -1
Omission rate: 0.045
Question difficulty: 0.6165

Question 41107 Average Human Rating: 3.2000 Average Human Difficulty: 0.2800

What hormone is lipid-soluble?

A. Anti-diuretic Hormone
B. Thyroid Hormone *correct answer*
C. Insulin
D. Calcitonin
E. Bony Congruence

Letter	High	Middle	Low	Total	Diff	Automated System	
A	2	2	2	6	0	0.079154	
B	31	51	24	106	7	0.080586	Correct
C	0	4	3	7	-3	0.0321398	
D	2	4	2	8	0	0.0531068	
E	0	0	0	0	0	0.0	
OMIT	0	1	5	6	0		
TOTAL	35	62	36	133	0		

Discriminating power: 0
Omission rate: 0.045
Question difficulty: 0.797

Question 41393 Average Human Rating: 3.0000 Average Human Difficulty: 0.1837

What protein is the component of thick filament within skeletal muscle?

A. Actin
B. Myosin *correct answer*
C. Troponin
D. Tropomyosin
E. Acetylcholine

Letter	High	Middle	Low	Total	Diff	Automated System	
A	2	2	3	7	-1	0.0560344	
B	31	55	27	113	4	0.0774978	Correct
C	1	0	1	2	0	0.0340632	
D	1	0	0	1	1	0.0435784	
E	0	2	0	2	0	0.0123006	
OMIT	0	3	5	8	0		
TOTAL	35	62	36	133	0		

Discriminating power: 1
Omission rate: 0.06
Question difficulty: 0.8496

Question 41354 Average Human Rating: 3.2321 Average Human Difficulty: 0.7321

What hormone has a negative influence on growth?

A. Thyroid Hormone
B. Insulin
C. Cortisol *correct answer*
D. Testosterone/Estrogen
E. ACTH

Letter	High	Middle	Low	Total	Diff	Automated System	
A	9	9	7	25	2	0.0839426	
B	4	11	1	16	3	0.0321398	
C	10	24	11	45	-1	0.0114472	Correct
D	3	2	1	6	2	0.0528876	
E	9	14	11	34	-2	0.0581802	
OMIT	0	2	5	7	0		
TOTAL	35	62	36	133	0		

Discriminating power: 1
Omission rate: 0.053
Question difficulty: 0.3383

Question 39695 Average Human Rating: 3.6667 Average Human Difficulty: 0.8571

Your hypothalamus has secreted some CRH (Corticotropic releasing hormone), where has this CRH gone?

A. Sympathetic Preganglionic Fibre
B. Posterior Pituitary Gland
C. Anterior Pituitary Gland *correct answer*
D. Adrenal Cortex
E. Adrenal Medulla

Letter	High	Middle	Low	Total	Diff	Automated System	
A	2	2	4	8	-2	0.012006	
B	0	1	1	2	-1	0.0642344	
C	21	38	16	75	5	0.0654924	Correct
D	10	16	5	31	5	0.0289488	
E	1	2	2	5	-1	0.0336982	
OMIT	1	3	8	12	0		
TOTAL	35	62	36	133	0		

Discriminating power: -1
Omission rate: 0.09
Question difficulty: 0.564

Question 41292 Average Human Rating: 3.2745 Average Human Difficulty: 0.3529

Which hormone directly effects the BMR?

A. Adrenaline
B. GH
C. Thyroid hormone *correct answer*
D. Cortisol
E. Calcitonin

Letter	High	Middle	Low	Total	Diff	Automated System	
A	0	0	1	1	-1	0.0178542	
B	2	0	4	6	-2	0.012119	
C	27	49	21	97	6	0.0819652	Correct
D	3	8	2	13	1	0.0097082	
E	3	2	4	9	-1	0.0575654	
OMIT	0	3	4	7	0		
TOTAL	35	62	36	133	0		

Discriminating power: -1
Omission rate: 0.053
Question difficulty: 0.7293

Question 41370 Average Human Rating: 2.9643 Average Human Difficulty: 0.2321

What am I? Regulates body temperature, water balance, sleep-cycle control, appetite, sexual arousal, pituitary and endocrine function?

A. Hypothalamus *correct answer*
B. Midbrain
C. Hydocephalus
D. Midbain
E. Pituitary Gland

Letter	High	Middle	Low	Total	Diff	Automated System	
A	33	58	29	120	4	0.0323932	Correct
B	0	0	0	0	0	0.0072606	
C	0	1	1	2	-1	0.004934	
D	0	0	0	0	0	0.0	
E	2	1	3	6	-1	0.0871852	
OMIT	0	2	3	5	0		
TOTAL	35	62	36	133	0		

Discriminating power: -1
Omission rate: 0.038
Question difficulty: 0.9023

Question 41409 Average Human Rating: 3.3115 Average Human Difficulty: 0.4754

What neural pathway will the motor neurons involved in chewing the steak follow?

A. Lateral Spinothalamic Pathway
B. Non Corticospinal Pathway *correct answer*
C. Rubrospinal Pathway
D. Corticospinal Pathway
E. Reticulospinal Pathway

Letter	High	Middle	Low	Total	Diff	Automated System	
A	1	3	1	5	0	0.085829	
B	23	40	18	81	5	0.090561	Correct
C	2	4	3	9	-1	0.0960282	
D	8	10	7	25	1	0.0819666	
E	1	1	3	5	-2	0.1020768	
OMIT	0	4	4	8	0		
TOTAL	35	62	36	133	0		

Discriminating power: 0
Omission rate: 0.06
Question difficulty: 0.609

Question 40946 Average Human Rating: 3.1296 Average Human Difficulty: 0.5185

If I were walking across campus to my lecture, which pathway would I be using?

A. Corticospinal
B. Medial Lemniscal Tract
C. Lateral Spinothalamic Tract
D. Pyrimidal Tract
E. Non-corticospinal *correct answer*

Letter	High	Middle	Low	Total	Diff	Automated System	
A	6	4	1	11	5	0.0043274	
B	1	4	2	7	-1	0.030261	
C	1	4	1	6	0	0.0039614	
D	3	5	3	11	0	0.0066804	
E	24	44	25	93	-1	0.00956	Correct
OMIT	0	1	4	5	0		
TOTAL	35	62	36	133	0		

Discriminating power: -1
Omission rate: 0.038
Question difficulty: 0.6992

Set:	2
Total number of students:	887
Total number of questions:	132
Percentage used of highest correlated students:	0.15
Percentage used of highest correlated questions:	0.25
Size of cohort:	133
Initial exam size:	33
New exam size:	20

Appendix I

Course 1 Results of Lucene Indexing and Searching

On the following page my be found an example output file for the Course 1 data from Lucene which shows one iteration of a matching run with the "hits" answer selection method.

From left to right the columns across the top of the document are: the question id, the course id, the experiment, the weight of the question (see Section 4.4.2 for an explanation of this process), the correct answer, the chosen answer, if the chosen answer is correct (1=correct, 0=incorrect), the variance, the standard deviation, the mean, the median, the minimum score, the maximum score, the range of scores, a distance measure for the scores and the number of hits of the correct answer option in the comparison files.

question	co	ex	weight	co	ch	is_	variance	stddev	mean	median	min_score	max_score	range_scor	dist_1_2	num_
41282	2	1	0.198	D	C	0	0.0020666	0.0454603	0.0331562	0.0122647	0.0058305	0.218642	0.212811	0.0803906	59
41332	2	1	0.211	D	A	0	0.0005567	0.0235942	0.0246395	0.0196919	0.0038725	0.136669	0.132797	0.0018172	135
41384	2	1	0.242	E	B	0	3.478E-05	0.0058972	0.013418	0.0102294	0.0078688	0.0289808	0.0211121	0.0029933	28
27535	2	1	0.255	B	B	1	0.0010959	0.0331048	0.0170774	0.0016082	0.0006964	0.367899	0.367202	0.204961	237
39887	2	1	0.257	B	B	1	0.0041287	0.0642552	0.0214138	0.006663	0.0047114	0.573236	0.568524	0.121785	187
43980	2	1	0.264	D	D	1	0.0205435	0.14333	0.13223	0.0920182	0.0209394	0.854304	0.833365	0.413639	56
41362	2	1	0.317	B	D	0	0.0020721	0.0455201	0.171584	0.148361	0.148361	0.282009	0.133648	0.119191	7
32648	2	1	0.328	E	A	0	0.0007552	0.0274806	0.0286509	0.0209087	0.0047005	0.142082	0.137382	0.0696643	46
43780	2	1	0.33	B	E	0	0.0017895	0.0423029	0.020628	0.007979	0.0012766	0.30496	0.303683	0.0626577	163
44100	2	1	0.346	D	A	0	0.000123	0.0110917	0.047948	0.0424798	0.0353998	0.073132	0.0377321	0.0098214	10
35088	2	1	0.35	E	E	1	0.0005123	0.0226341	0.0234446	0.017303	0.0099899	0.145875	0.135885	0.0694701	43
31522	2	1	0.352	D	D	1	0.0004089	0.020222	0.0375193	0.0343037	0.0207912	0.138268	0.117477	0.0345671	47
41356	2	1	0.378	C	C	1	0.0274157	0.165577	0.192854	0.135276	0.0307828	0.64782	0.617037	0.104569	53
41354	2	1	0.382	C	D	0	0.0018476	0.0429841	0.0498846	0.0307339	0.0186276	0.174815	0.156188	0.0035841	54
39542	2	1	0.389	A	C	0	0.0001913	0.0138324	0.0689094	0.066942	0.0576792	0.0956617	0.0379825	0.0009914	11
41357	2	1	0.404	D	B	0	0.001912	0.0437262	0.0544521	0.0386118	0.0180619	0.203839	0.185777	0.0101237	64
44103	2	1	0.406	D	D	1	0.0026496	0.051474	0.0291845	0.0090765	0.0055012	0.274226	0.268725	0.0481458	176
28890	2	1	0.41	D	C	0	0.0032008	0.0565753	0.0580957	0.0409442	0.0248159	0.346539	0.321723	0.0858386	65
27641	2	1	0.419	C	B	0	0.0002306	0.0151843	0.0131555	0.0061802	0.0007341	0.0594738	0.0587398	0.0015666	137
43485	2	1	0.425	E	D	0	0.0271112	0.164655	0.104159	0.0479528	0.0424141	0.768704	0.72629	0.0948475	34
39594	2	1	0.432	C	B	0	0	0	0.379226	0.379226	0.379226	0.379226	0	0	2
44010	2	1	0.447	D	B	0	0.0260845	0.161507	0.124578	0.0392202	0.0280144	0.738236	0.710221	0.217964	35
30729	2	1	0.451	B	E	0	0.0012247	0.0349958	0.0146776	0.001704	0.0005443	0.322238	0.321694	0.0635139	213
44128	2	1	0.454	B	D	0	0.0055038	0.0741875	0.070988	0.0518209	0.0314347	0.446023	0.414588	0.315629	30
28037	2	1	0.46	C	C	1	0.0170481	0.130568	0.151203	0.0902076	0.064434	0.563776	0.499342	0.0747305	31
43480	2	1	0.461	C	C	1	0.0152685	0.123566	0.120068	0.101546	0.0190945	0.599409	0.580315	0.123403	52
31982	2	1	0.468	B	D	0	0.0006303	0.0251048	0.0153817	0.0105564	0.0063981	0.23599	0.229592	0.0839335	163
40863	2	1	0.495	B	B	1	0.0248487	0.157635	0.169527	0.0944653	0.0687634	0.801411	0.732647	0.34216	32
41436	2	1	0.498	B	B	1	0.0006558	0.025609	0.0325183	0.0248703	0.014655	0.14196	0.127305	0.0386072	39
31281	2	1	0.512	C	C	1	0.0065946	0.0812072	0.0618093	0.0231576	0.0081035	0.49866	0.490557	0.184491	95
43631	2	1	0.513	B	A	0	0.0007573	0.0275187	0.0198534	0.0081532	0.0049416	0.126216	0.121275	0.0032329	111
32440	2	1	0.515	A	D	0	0.0083878	0.0915847	0.0704218	0.045018	0.0186875	0.502858	0.484171	0.0538107	61
39611	2	1	0.539	B	B	1	0.0025845	0.0508377	0.046037	0.0291121	0.0171545	0.250828	0.233673	0.0032166	45
41481	2	1	0.548	E	E	1	0.0707882	0.266061	0.199968	0.0757297	0.0421968	1.03739	0.995192	0.148199	31
32030	2	1	0.549	C	C	1	0.0046874	0.0684648	0.0336334	0.0091115	0.004602	0.484428	0.479826	0.069204	144
27482	2	1	0.552	E	E	1	0.0002917	0.0170789	0.0880514	0.0898684	0.0573339	0.139027	0.0816933	0.0292835	31
40973	2	1	0.552	B	A	0	0.0009565	0.030927	0.0394876	0.0216014	0.0086117	0.120939	0.112328	0.0047063	41
41589	2	1	0.567	B	A	0	0.0018006	0.0424332	0.0592778	0.0449856	0.0229076	0.231727	0.20882	0.0886118	31
41366	2	1	0.57	D	D	1	0.0036911	0.0607544	0.0597225	0.0406119	0.0101387	0.269028	0.258889	0.0351091	94
38353	2	1	0.587	C	A	0	0.000184	0.0135647	0.0501189	0.0535359	0.027895	0.0658903	0.0379953	0	12
39695	2	1	0.591	C	C	1	9.79E-05	0.0098946	0.0087484	0.0053409	0.0026478	0.0613511	0.0587032	0.0216828	72
41323	2	1	0.597	D	E	0	0.0003045	0.0174505	0.0123879	0.0064288	0.0028938	0.104346	0.101452	0.0488147	44
41409	2	1	0.604	B	E	0	0.0013555	0.0368175	0.0212308	0.0036266	0.0021999	0.199696	0.197496	0.0243544	192
44127	2	1	0.607	E	E	1	0.0007907	0.0281193	0.0182392	0.0041991	0.0024244	0.144432	0.142008	0.0153089	175
43372	2	1	0.618	B	B	1	0.0017717	0.042091	0.0679172	0.0564302	0.0353272	0.263279	0.227952	0.111952	32
30718	2	1	0.619	A	E	0	0.0052408	0.0723934	0.0559845	0.0317796	0.0213335	0.321244	0.29991	0.0458919	29
43579	2	1	0.62	D	C	0	0.0421195	0.20523	0.242543	0.169856	0.127392	0.810055	0.682663	0.552039	9
41387	2	1	0.63	A	A	1	0.0018653	0.0431889	0.202956	0.19922	0.11502	0.271104	0.156084	0.0054774	10
41288	2	1	0.635	E	B	0	0.0017156	0.0414195	0.0549885	0.0432753	0.0284694	0.227932	0.199463	0.0797127	26
31652	2	1	0.654	A	C	0	0.0035247	0.0593688	0.0378277	0.0130759	0.0062328	0.409767	0.403534	0.166292	119
44056	2	1	0.659	A	B	0	0.0003022	0.0173836	0.016398	0.0106066	0.0031883	0.10767	0.104482	0	125
44001	2	1	0.67	D	C	0	0.0010113	0.0318012	0.0288507	0.0187981	0.0029403	0.188316	0.185375	0.0272215	150
43999	2	1	0.67	D	D	1	0.0004853	0.0220298	0.0189573	0.0119797	0.0060506	0.138794	0.132744	0.0844476	42
38314	2	1	0.675	B		0	0	0	0	0	0	0	0	0	0
43599	2	1	0.675	E	C	0	0.0006797	0.0260706	0.0225759	0.01303	0.0042919	0.116504	0.112212	0.0002195	84
40946	2	1	0.676	E	C	0	0.0067571	0.0822014	0.061934	0.0299316	0.0217784	0.441719	0.419941	0.106361	43
40328	2	1	0.681	D	C	0	0.0075364	0.0868127	0.0759558	0.0426275	0.0073161	0.444055	0.436739	0.0557525	74
41443	2	1	0.681	A	D	0	0.0001856	0.013624	0.0126504	0.006462	0.0014663	0.0765647	0.0750984	0.0122732	129
40435	2	1	0.688	B	B	1	0.0033768	0.05811	0.0459122	0.0211927	0.0107039	0.308951	0.298247	0.100204	87
40329	2	1	0.691	D	D	1	0.0138317	0.117608	0.0667066	0.0270384	0.0047511	1.0228601	1.0181	0.407125	166
38964	2	1	0.693	D	A	0	0.0010629	0.0326017	0.0140241	0.0036097	0.0018048	0.270553	0.268748	0.114288	157

Figure I.1: An example output file from the Lucene runs.

Bibliography

[1] Sarah K. K. Luger and Jeff Bowles. Two methods for measuring question difficulty and discrimination in incomplete crowdsourced data. In *Proceedings of The First AAAI Conference on Human Computation and Crowdsourcing (HCOMP-13), Palm Springs, CA, USA.*, 2013.

[2] Norman E. Gronlund. *Measurement and Evaluation in Teaching. 4th ed.,*. Macmillan, 1981.

[3] Alejandro Figueroa and John Atkinson. Using dependency paths for answering definition questions on the web. In *WEBIST*, 2009.

[4] Silvia Quarteroni. *Advanced Techniques for Personalized, Interactive Question Answering*. PhD thesis, York University, 2007.

[5] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *In Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004*, 2004.

[6] John Prager. *Open-Domain Question-Answering. Foundations and Trends in Information Retrieval*. now Publishers, Inc, 2006.

[7] Hyo-Jung Oh, Chung-Hee Lee, Hyeon-Jin Kim, and Mjung-Gil Jang. Descriptive question answering in encyclopedia. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 2005.

[8] Richard Phelps. Are us students the most heavily tested on earth? *Educational Measurement: Issues and Practice*, 15(3), 1996.

[9] Google define mercury. `http://www.google.com/search?hl=en&q=define:+mercury&btnG=Search`, December 13, 2010.

[10] Paul Denny. Peerwise. `http://PeerWise.cs.auckland.ac.nz/`.

[11] Andrew M. Olney, Arthur Graesser, and Natalie K Person. Question generation from concept maps. *Dialogue and Discourse 3(2), 75-99*, 2012.

[12] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science, Volume 41, Issue 6, pages 391-407, September*, 1990.

[13] Thomas K. Landauer and Susan T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, Volume 104, No.2, pages 211-240*, 1997.

[14] Wikipedia. Regents exam. `http://en.wikipedia.org/wiki/Regents_exam`, 2010.

[15] Wikipedia. Multiple choice, 2010.

[16] Meredith D. Gall. The use of questions in teaching. *Review of Educational Research*, 1970.

[17] Tamar Lewin. Harvard and m.i.t. team up to offer free online courses. *The New York Times*, May 2, 2012.

[18] Michael Winerip. Facing a robo-grader? just keep obfuscating mellifluously. *The New York Times*, April 22, 2012.

[19] Quora. Quora. `http://www.quora.com`, 2013.

[20] Piazza company website pooja sankar. `http://www.piazza.com`, December 30, 2013.

[21] Paul Denny, Andrew Luxton-Reilly, and John Hamer. Student use of the peerwise system. In *ITiSCE '08: Proceedings of the 13th annual conference on Innovation and technology in computer science education Pages 73-77*, 2008.

[22] Andrew Ng and Daphne Koller. Cousera. `https://www.coursera.org/`, 2013.

[23] Edx online education platform mit and harvard. `http://www.edx.org/`, December 30, 2013.

[24] Yale university's open yale online education portal open yale. http://oyc.yale.edu/, December 30, 2013.

[25] Tamar Lewin. U.s. teams up with operator of online courses to plan a global network. http://www.nytimes.com/2013/11/01/education/us-plans-global-network-of-free-online-courses.html?, October 31, 2013.

[26] Norman E. Gronlund. *Preparing Criterion-Referenced Tests for Classroom Instruction*. Macmillan, 1973.

[27] K. Barker, V. K. Chaudhri, S. Y. Chaw, P.E. Clark, J. Fan, D. Israel, S. Mishra, B. Porter, P. Romero, D. Tecuci, and P. Yeh. A question-answering system for ap chemistry: Assessing kr technologies. *KR*, 2004.

[28] Stephen Bottemly and Paul Denny. A participatory learning approach to biochemistry using student authored and evaluated multiple-choice questions. *Biochemistry and Molecular Biology Education, Volume 39, No. 5, pages 352-361*, 2011.

[29] Vinay K. Chaudhri, Britte Haugan Cheng, Adam Overholtzer, Jeremy Roschelle, Aaron Spaulding, Peter Clark, Mark Greaves, and Dave Gunning. Inquire biology: A textbook that answers questions. *AI Magazine, Volume 34, No.3, Fall 2013, pages 55-72*, 2013.

[30] Paul Denny. The effect of virtual achievements on student engagement. In *CHI'13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Pages 763-772*, 2013.

[31] Michael Bieber, Jia Shen, Denzhi Wu, and S. Roxanne Hiltz. Participatory learning approach. *Encyclopedia of Distance Learning, Volume 3, ICI Global: Hershey, PA, USA pages 1467-1472*, 2005.

[32] Benedict Carey. Frequent tests can enhance college learning, study finds. http://www.nytimes.com/2013/11/21/education/frequent-tests-can-enhance-college-learning-study-finds.html?_r=0, November 21, 2013.

[33] Peder J. Johnson, Timothy E. Goldsmith, and Kathleen W. Teague. Locus of the predictive advantage in pathfinder-based representations of classroom knowledge. *Journal of Educational Psychology, 86(4) Pages 617-626*, 1994.

[34] A.R. Merchant and K.W. McGregor. Reflections on using student-authored questions to encourage learning in physics. In *Proceedings of the blended learning in science teaching and learning symposium, The University of Sydney*, 2005.

[35] Timothy E. Goldsmith, Peder J. Johnson, and William H. Acton. Assessing structural knowledge. *Journal of Educational Psychology, 83(1) Pages 88-96*, 1991.

[36] Norman E. Gronlund. *Constructing Achievement Tests*. Prentice Hall, 1977.

[37] Wikipedia. Item response theory, 2011.

[38] A. A. Beguin and C. A. W. Glas. Mcmc estimation and some model-fit analysis of multidimensional irt models. *Psychometrika, Volume 66, No.4, pages 541-562*, 2001.

[39] Richard J. Patz and Brian W. Junker. Applications and extensions of mcmc in irt: Multiple item types, missing data and rated responses. *Journal of Educational and Behavioral Statistics, Volume 24, No.4, Winter 1999, pages 342-366*, 1999.

[40] Tiffany Barnes. The q-matrix method: Mining student response data for knowledge. In *Proceedings of AAAI 2005 Educational Data Mining Workshop*, 2005.

[41] Klaus D. Kubinger and Christian H. Gottschall. Item difficulty of multiple choice tests dependent on different item response formats – an experiment in fundamental research on psychological assessment. *Psychology Science, 49(4) Pages 361-374*, 2007.

[42] Brooke Soden Hensler and Joseph Beck. Better student assessing by finding difficulty factors in a fully automated comprehension measure. *Intelligent Tutoring Systems Pages 21-30*, 2006.

[43] Daniel Jurafsky and Lames H. Martin. *Speech and Language Processing*. Pearson Prentice Hall, 2008.

[44] M. R. Quillian. *Word concepts: A theory and simulation of some basic semantic capabilities*. Brachman and Levesque (1985), 1967.

[45] R. C. Schank and R. Abelson. *Scripts, Plans, Goals, and Understanding*. Erlbaum, Hillsdale, N.J., 1977.

[46] John F. Sowa. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, 1984.

[47] Charles J. Fillmore. The case for case. *Bach and Harms*, 1968.

[48] Framenet website. `https://framenet.icsi.berkeley.edu/fndrupal/about`, 2014.

[49] ITiCSE '09 Proceedings of the 14th annual ACM SIGCSE conference on Innovation and technology in computer science education, Pages 11-15. *Coverage of course topics in a student generated MCQ repository*, 2009.

[50] Princeton University. "about wordnet.". `http://wordnet.princeton.edu/`, 2010.

[51] Claudia Leacock, George A. Miller, and Martin Chodorow. Using corpus statistics and wordnet relations for sense identification. In *Computational Linguistics 24(1)*, 1998.

[52] Ruslan Mitkov and Le An Ha. Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing, Edmonton, Canada, Pages 17-22*, May 2003.

[53] Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, 1998.

[54] Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, (2010), Volume 37, Pages 141-188*, 2010.

[55] Apache. Lucene. `https://lucene.apache.org/`, 2013.

[56] Michael McCandless, Eric Hatcher, and Otis Gospodnetic. *Lucene in Action, Second Edition*. Manning, 2010.

[57] Toefl question set with answers used in lsa research and made available by prof. lindauer at the university of colorado at boulder, lsa and nlp research labs via email correspondence. thomas k. lindauer. `http://lsa.colorado.edu/`, September 25, 2013.

[58] Acl's toefl synonym questions (state of the art) association of computational linguistics web. `http://aclweb.org/aclwiki/index.php?title=TOEFL_Synonym_Questions_(State_of_the_art)`, December 30, 2013.

[59] Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*, 1997.

[60] Michael Poprat, Elena Beisswanger, and Udo Hahn. Building a biowordnet by using wordnet's data formats and wordnet's software infrastructure – a failure story. In *Software Engineering Testing and Quality Assurance for Natural Language Processing*, 2008.

[61] Ana Licuanan Jinxi Xu and Ralph Weischedel. Trec2003 qa at bbn: Answering definitional questions. In *The Twelfth Text Retrieval Conference (TREC 2003)*, 2004.

[62] Le An Ha. Multiple choice test item generation: A demo. In *RANLP-07 Workshop: Computer-Aided Language Processing (CALP'07)*, 2007.

[63] Robert Michael Foster. Improve the output from a mcq test item generator using statistical nlp. In *In proceedings of the International Conference in Alternative Learning Technologies (ICALT)*, 2010.

[64] Ruslan Mitkov, Le An Ha, Andrea Varga, and Luz Rello. Semantic similarity of distractors in multiple-choice tests: Extrinsic evaluation. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, 2009.

[65] Ruslan Mitkov, Le An Ha, and Nikiforos Karamanis. A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering 12(2). 177-194*, 2006.

[66] Miri Barak and Sheizaf Rafaeli. On-line question-posing and peer-assessment as means for web-based knowledge sharing in learning. *International Journal of Human-Computer Studies*, 2004.

[67] European Conference on IR. *Web-Based Multiple Choice Question Answering for English and Arabic Questions*, 2006.

[68] Alain Lifchitz, Sandra Jhean-Larose, and Guy Denhiere. Effect of tuned parameters on a lsa multiple choice questions answering model. *Behavior Research Methods*, 2009.

[69] Robert Munro and Harry Tily. The start of the art: An introduction to crowdsourcing technologies for language and cognition studies. In *Workshop on Crowdsourcing Technologies for Language and Cognition Studies.*, 2011.

[70] Amazon's mechanical turk website amazon. `http://www.mturk.com`, December 30, 2013.

[71] Crowdflower company website lukas biewald and chris van pelt. `http://www.crowdflower.com`, December 30, 2013.

[72] Chris Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *EMNLP 09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 2009.

[73] Tanja Janssen. Instruction in self-questioning as a literary reading strategy: An exploration of empirical research. *F1-Educational Studies in Language and Literature, Volume 2, Number 2, 95-120*, 2002.

[74] Jane McGonigal. *Reality is broken: Why Games Make Us Better and How They Can Change the World*. Penguin Press HC, 2011.

[75] Sarah K. K. Luger and Jeff Bowles. 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, DUBMOD at CIKM 2013 San Francisco, CA, USA, October 28, 2013. Pages 29-32.*, 2013.

[76] Frederick B. Davis. *Educational Measurements and their Interpretation*. Wadsworth Publishing Company, Inc., 1964.

[77] Ellen M Voorhees. Overview of the trec 2004 question answering track. In *The Thirteenth Text REtrieval Conference (TREC 2004)*, 2004.

[78] Ellen M Voorhees and Hoa Trang Dang. Overview of the trec 2005 question answering track. In *The Fourteenth Text REtrieval Conference (TREC 2005)*, 2005.

[79] Ellen M Voorhees. Overview of the trec 2003 question answering track. In *The Twelfth Text Retrieval Conference (TREC 2003)*, 2004.

[80] Frank Keller, Maria Lapata, and Olga Ourioupina. Using the web to overcome data sparseness. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002.

[81] Stephanie M. Zinn. *McGraw-Hill's SAT Subject Test: Biology E/M*. McGraw-Hill, 2009.

[82] Gabrielle I. Edwards, Marion Cimmino, Frank J. Foder, and G. Scott Hunter. *Barron's Regents Exams and Answers. Biology: The Living Environment*. Barron's Educational Series, Inc., New York, 2008.

[83] Deborah T. Goldberg. *Barron's SAT Subject Test Biology E/M*. Barron's Educational Series, Inc., 2009.

[84] Laurie Ann Callihan. *The Best Test Preparation for the CLEP Biology Exam with TESTware on CD-ROM*. Research and Education Association, Inc, New York, 2008.

[85] Phillip E. Pack. *CliffsAP 5 Biology Practice Exams*. Wiley Publishing, Inc., 2006.

[86] Alison Pitt. *Roadmap to the Regents: Living Environment*. Princeton Review Publishing, 2003.

[87] Roy Thomas Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, 2000.

[88] Norman Walsh. Marklogic database server. `http://www.marklogic.com/`, 2010.

[89] Norman Walsh. Xml calabash. `http://www.xmlcalabash.com`, 2010.

[90] Norman Walsh, Alex Milowski, and Henry Thompson. Xproc. `http://www.w3.org/TR/xproc/`, 2010.

[91] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004.

[92] Alfred V. Aho, John E. Hopefroft, and Jeffrey D. Ullman. *Design and Analysis of Computer Algorithms*. Addison-Wesley Publishing Company, Inc., 1974.

[93] Michael Sipser. *Introduction to the Theory of Computation*. Course Technology, 2005.

[94] Adrian Bondy and U.S.R. Murty. *Graph Theory (Graduate Texts in Mathematics)*. Springer, 2010.

[95] William Feller. *An Introduction to Probability Theory and Its Applications*. John Wiley and Sons, Inc., 1950.

[96] Sarah Luger. A graph theory approach for generating multiple choice exams. In *2011 AAAI Fall Symposium on Question Generation*, 2011.

[97] Robert Munro, Schuyler E. Erle, and Tyler Schnoebelen. Analysis after action report for the crowdsourced aerial imagery assessment following hurricane sandy. In *10th International Conference on Information Systems for Crisis Response and Management. Baden, Baden Germany*, 2013.

[98] Jennifer Chu-Carroll and James Fan. Leveraging wikipedia characteristics for search and candidate generation. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence, Palo Alto, CA, USA*, 2011.

[99] Christof Muller and Iryna Gurevych. Using wikipedia and wiktionary in domanspecific information retrieval. In *CLEF '08: Proceedings of the 9th Cross-language evaluation forum conference on evaluating systems for multilingual and multimodal information access. Pages 219-226*, 2008.

[100] Rakesh Agrawal, Sreenivas Gollapudi, Anitha Kannan, and Krishnaram Kenthapadi. Data mining for improving textbooks. *SIGKDD Explorations, Volume 13, Number 2, pages 7-10.*, 2011.

[101] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John

Prager, Nico Schlaefer, and Chris Welty. Building watson: An overview of the deepqa project. In *AI Magazine*, 2011.

[102] David Ahn, Valentin Jijkoun, Gilad Mishne, Karin Muller, Maarten de Rijke, and Stefan Schlobach. Using wikipedia at the trec qa track. text retrieval conference. using wikipedia at the trec qa track. In *TREC 2004*, 2004.

[103] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A core of semantic knowledge unifying wordnet and wikipedia. In *International World Wide Web Conference, Track: Semantic Web, Session: Ontologies. IW3C2*, 2007.

[104] Combinatorics, Math.co. *Finding bipartite subgraphs efficiently*, 15 May, 2009.

[105] Gabriela Alexe, Sorin Alexe, Yves Crama, Stephan Foldes, Peter L. Hammer, and Bruno Simeone. Consensus algorithms for the generation for all maximal bicliques. *Discrete Applied Mathematics (145), 11-21. Graph Optimization IV*, 2004.